

A Thesis Submitted for the Degree of PhD at the University of Warwick

Permanent WRAP URL:

<http://wrap.warwick.ac.uk/160610>

Copyright and reuse:

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it.

Our policy information is available from the repository home page.

For more information, please contact the WRAP Team at: wrap@warwick.ac.uk

Statistical modelling of spatio-temporal decision data

by

Lara Vomfell

A thesis submitted in partial fulfilment of the requirements for the

degree of

Doctor of Philosophy

University of Warwick, Warwick Business School

March 2021

Contents

Acknowledgments	vii
Declaration	viii
Abstract	ix
Abbreviations	x
Introduction	1
Ethnic bias in stop and search	2
Modelling domestic abuse data	5
Robust inference under model misspecification	9
1 Officer bias, over-patrolling, and ethnic disparities in stop and search	13
1.1 Introduction	13
1.2 Related work	15
1.3 Data	18
1.4 Methods	19
1.4.1 Measures of over-searching	21
1.4.2 Multinomial model	21
1.5 Results	24
1.5.1 Inference of search shares	25
1.5.2 Measures of officer over-searching	27
1.5.3 Officer- compared to force-level bias	28
1.6 Discussion	32
Appendices	
1.A Sample selection	35
1.B Additional model results	36

1.C	AR(1) model	37
2	(No) Spillovers in reporting domestic abuse to police	42
2.1	Introduction	42
2.2	Motivation	45
2.2.1	Spillovers in crime	45
2.2.2	Reporting domestic abuse	48
2.2.3	Spillover channels	51
2.3	Data	53
2.4	Method	55
2.4.1	Model	55
2.4.2	Inference	58
2.5	Results	65
2.6	Discussion	70
	Appendices	
2.A	Likelihood	72
2.B	Inference algorithm	73
3	Robust Bayesian Inference for Discrete Outcomes with the Total Variation Distance	75
3.1	Introduction	75
3.2	Divergences & Inference	78
3.3	Motivation	79
3.4	Estimating the Total Variation Distance (TVD)	81
3.4.1	Exponential Concentration Inequalities	82
3.4.2	Almost sure convergence	84
3.4.3	Provable robustness	84
3.4.4	Consistency	85
3.5	Nonparametric Learning	85
3.6	Experiments	88
3.6.1	Evaluation criteria	89
3.6.2	ε -contamination	90
3.6.3	Zero-inflation	91
3.6.4	Probit	92
3.6.5	Neural Network	92
3.6.6	Modelling crime incidence	93
3.7	Conclusion	93

Appendices	
3.A Proofs	94
3.B Full experimental results	102
Conclusion	105
Summary	105
Discussion	109
Bibliography	115

List of Tables

1.1	Means, standard deviations (SD), minima and maxima of variables used in the multinomial model in Section 1.4.2.	20
1.2	Details of matching and exclusion criteria applied to incidents , crimes , stops and officers	36
1.3	Estimates and 90% uncertainty intervals (UI) for model parameters in Equation (1.2).	37
3.1	Overview of statistical quantities used in Chapter 3.	77

List of Figures

1.1	Posterior densities of search shares p_{ite}	25
1.2	Posterior densities of the coefficients used to infer search shares p	26
1.3	Posterior densities of O_{ite}^S and O_{ite}^P	29
1.4	Histograms of the posterior probabilities of O_{ite}^S and O_{ite}^P above 1	29
1.5	Decomposition of over-searching	31
1.6	Disaggregated densities of posterior distributions of search shares	37
1.7	Posterior densities of O^S and O^P disaggregated by time	38
1.8	Comparison of observed search counts to predicted search counts	38
1.9	Densities of AR(1) coefficients for time series of O^S and O^P	40
2.1	Time and location occurrence of events	55
2.2	Estimated background components	67
2.3	Estimated triggering components	68
2.4	Share of households in neighbourhood living in non-detached houses against the mean number of reports triggered.	69
2.5	Deviation of the transformed time sequence (purple) from the the- oretical (black) sequence with 95% confidence bands (grey).	70
3.1	Discrete count data with some outliers (grey histogram) are mod- elled with a Poisson distribution.	78
3.2	Difference in inference outcomes for the ε -contamination model of Equation (3.9) between using the TVD and the KLD as k is varied and $\varepsilon = 0.15$. Left : Absolute difference between inferred and true value of λ ; Middle : Absolute out-of-sample prediction error; Right : Predictive likelihood on out-of-sample data.	90

3.3	Difference in inference outcomes for the zero-inflation model between using the TVD and the KLD as the proportion ε of zeros is varied. Left: Absolute difference between inferred and true value of β ; Middle: Absolute out-of-sample prediction error; Right: Predictive likelihood on out-of-sample data.	91
3.4	Predictive likelihoods for the Probit models (top) and single-layer Neural Networks (bottom).	92
3.5	Incidence of sexual offences and model inference.	94
3.6	Full simulation results comparing inference using the Kullback-Leibler Divergence (KLD), the Total Variation Distance (TVD) and a fully Bayesian approach.	103
3.7	Predictive accuracy from 50 random splits for Probit models (top) and Neural Networks (bottom).	104

Acknowledgments

I would like to thank my supervisor Neil Stewart and so many people at Warwick for their support and advice. I am grateful for the financial support from the Bridges Leverhulme Programme that opened many doors for me.

Mama, thank you for the mutual tethering to earth and your unwavering confidence. To my friends and chayas, thank you for entertaining my pursuit of conventional accolades. Jeremias, without you none of this would have happened.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. All work presented was carried out in collaboration with co-authors as follows:

Chapter 1 is co-authored with Neil Stewart (Warwick Business School, University of Warwick). Data was provided by West Midlands Police to Neil Stewart. I developed the concept of the paper, collated and analysed the data. I wrote the manuscript with edits by Neil Stewart. The paper is published at *Nature Human Behaviour*.

Chapter 2 is co-authored with Jan Povala (Department of Mathematics, Imperial College London). Data was provided by an unnamed English police force. I developed the concept of the paper, collated and analysed the data. Jan Povala and I developed the extension of the Zhuang and Mateu (2019) specification of the Hawkes process model together. I am responsible for the technical implementation. I wrote the manuscript with edits by Jan Povala.

Chapter 3 is co-authored with Jeremias Knoblauch (Department of Statistics, University of Warwick). Jeremias Knoblauch and I developed the concept of the paper together. Jeremias Knoblauch is responsible for the theoretical results (Propositions 2, 3 and 4 and Corollaries 1 and 2). Jeremias Knoblauch and I are responsible for the technical implementation of Bayesian Nonparametric Learning. I am responsible for the technical implementation and analysis of the experimental results. Data for the crime example was provided by an unnamed English police force and was analysed by me. Jeremias Knoblauch and I wrote the paper together. It is currently under review at the International Conference on Machine Learning.

Abstract

This thesis consists of three independent research studies in the fields of statistical and behavioural science. Each study is concerned with modelling complex spatio-temporal decisions recorded in police data. Analysing decisions at a high resolution requires a comprehensive understanding of the social phenomenon and data-generating mechanism, combined with careful modelling choices.

Chapter 1 is a novel model of ethnic bias at the officer-level in stop and search. Using a Bayesian hierarchical model, we model officer over-searching against two officer-specific baselines: the crime suspects that the officer encounters and the local patrolling area of the officer. We find that most police officers are biased against Black and Asian people in their search decisions, independently of which baseline we use. Furthermore, we decompose bias against ethnic minority groups into bias due to officer over-searching and over-patrolling.

Chapter 2 showcases the use of a spatio-temporal Hawkes-type point process to model the reporting of domestic abuse. Extending existing Hawkes models, we test for the existence of two spillover channels in crime victim reporting. Despite well-documented spillover effects in other human behaviour, we find no evidence to support such effects in the reporting of domestic abuse.

Chapter 3 introduces a new, robust statistical inference procedure for discrete outcomes. We propose using the Total Variation Distance together with Bayesian Nonparametric Learning to robustify inference. We show that this procedure possesses a range of desirable theoretical properties. Furthermore, we demonstrate that our method outperforms standard inference both in terms of inference and out-of-sample performance on simulated data. Lastly, we show that robust inference is important for modelling police-recorded incidence of sexual offences where fluctuations in reporting can drastically affect inference.

I conclude by discussing the importance of sophisticated statistical approaches to reflect often complicated underlying social phenomenon and the equally complex process by which it is recorded in data.

Abbreviations

BFGS Broyden–Fletcher–Goldfarb–Shanno algorithm

DP Dirichlet Process

KDE kernel density estimation

KLD Kullback-Leibler Divergence

NPL Bayesian Nonparametric Learning

ONS Office for National Statistics

TVD Total Variation Distance

Introduction

With the advent of large-scale and fine-grained behavioural data, behavioural science is facing a unique opportunity for expanding the boundaries of understanding human behaviour. However, this opportunity can only be fully realised if accompanied by sound methods.

Already, behavioural science has shifted tremendously by transitioning from predominantly lab-based, experimental studies to studying human decision-making in the field, too (DellaVigna, 2009; Maner, 2016). The increased availability and utilization of data on choices made in the real world have been transformative (Buyalskaya, Gallo, and Camerer, 2021). Now, behavioural science is tasked with the next step to advancing our discipline even further: A further sophistication of statistical literacy, modelling and application.

This PhD thesis highlights the enormous potential inherent in this endeavour. My work lies at the intersection of the behavioural and statistical sciences and seeks to advance our understanding of decision-making in the context of policing and crime. In the process, I address and resolve issues that arise from modelling ever more detailed data.

Decision-level data from police officers, crime perpetrators and victims are an exciting new source for understanding human behaviour. Cities, police departments and researchers alike are making their data openly available (e.g., Stanford Open Policing Project, 2021). This increase in availability of large-scale data has been accompanied by a proliferation in the use of sophisticated statistical and machine learning models.

At the same time, these developments are not without drawbacks. More fine-grained data require a careful and comprehensive interrogation of the entire context in which they emerged. Crime and police are central public issues that touch many people's lives. It is imperative that scientists are mindful of the complexities of working in this area: For some, police contact is a welcome response to crime victimization, for others it is a stigmatizing experience. In such

an intricate context full of ethical considerations, it is imperative that research is conducted to the highest standards of science (Bartlett, 2019). At the same time, it is these very complexities that make research in this area challenging. The chapters in this thesis highlight the importance of interrogating the entire context and provide methods and examples of how our understanding of a problem depends on the techniques used to approach it: Chapter 1 explores this in the context of ethnic bias in stop and search decisions, Chapter 2 in the context of investigations into the dynamics of crime reporting and Chapter 3 presents a generic robustness method to model misspecification for discrete outcomes.

Ethnic bias in stop and search

Stop and search is a central police power that allows police officers to stop individuals and search their belongings if the officers believe the person is carrying contraband or has committed a crime. While police consider stop and search a key tool to investigating and preventing crime, it is one of the more controversial policing practices. Stop and search rates exhibit persistent ethnic disparities, not just in the United Kingdom but in many countries (Committee on the Elimination of Racial Discrimination, 2015; Barnes, 2019; Human Rights Watch, 2020; Pierson et al., 2020). In the United Kingdom, Black and Asian people are between 3 and 9 times more likely to be stopped and searched by police than White people (Home Office, 2018b). Such disparities entail disproportionate contact with police for young Black and Asian men, not just at the initial search: A search can lead to feedback loops resulting in repeat contact with police (Kohler-Hausmann, 2013; Quinton, 2011; Sharp and Atherton, 2007). Furthermore, search encounters are invasive experiences which can result in significant psychological harm (Skogan, 2006; Geller et al., 2014; Del Toro et al., 2019).

Most standard approaches to studying ethnic bias in policing are rooted in the economics literature and seek to distinguish between the two dominant explanations for ethnic biases dominant in that field: statistical discrimination and prejudice. Statistical discrimination refers to the phenomenon where unbiased officers search ethnic minorities at higher rates but hit rates—the rate at which illegal items are recovered—at the margin for all ethnic groups are equal. The underlying idea is that some ethnic groups are more likely to carry illegal items and unbiased officers are only responding to such group-based differences. In contrast, ethnic prejudice refers to police officers who hold unfounded beliefs or

stereotypes about other ethnicities. This prejudice can be implicit or explicit and means that officers will continue to search ethnic minorities such that the marginal hit rate of minorities is lower than the marginal hit rate of White people. Both explanations, ethnic prejudice and statistical discrimination, predict the same outcome: the over-representation of some ethnicities in stop and search (Charles and Guryan, 2011). Consequently, much work has focused on deriving precise tests to distinguish between them.

These tests are called outcome tests: If officers are prejudiced against ethnic minorities, they will continue to search minority members in equilibrium even when the corresponding marginal success rate is lower than that of White people (Becker, 1957; Knowles, Persico, and Todd, 2001). In practice however, the unobserved nature of hit rates at the margins makes it difficult to apply this test and can only be done with strong assumptions (Ayres, 2002; Dharmapala and Ross, 2004; Engel and Tillyer, 2008). This fundamental issue with a marginal hit rate approach has recently been addressed in the form of threshold tests (Goel, Rao, and Shroff, 2016; Simoiu, Corbett-Davies, and Goel, 2017). These tests estimate the ethnicity-specific distribution of the likelihood to carry illegal items. In addition, they estimate the marginal threshold used by police to search the marginal ethnicity member. Lower thresholds for some ethnicities are then an indication of discrimination.

Even so, there remain three fundamental issues with these approaches. The first is pooling. The typical study of ethnic bias in policing activities analyses search counts pooled across an entire police force, effectively treating the police force as a homogeneous object of study. Recent work by Ross, Winterhalder, and McElreath (2018) demonstrates that even in a setting where all police officers are assumed to be biased against an ethnic group by construction, an analysis where data is pooled across officers can fail to detect this bias. This finding echoes Simpson’s paradox: Statistical relationships can be obscured or even reversed when one chooses an inappropriate level of analysis. Choosing the appropriate level of analysis is crucial and has immense consequences for the conclusions. We discuss this point in more detail in Chapter 1.

The second issue is the assumption of homogeneous behaviour within a police force. This issue is intimately related to pooling but worth addressing in more detail. There are (at least) two sources of variation between police officers: The first is different exposures and professional responsibilities which leads to variation between officers in who they interact with. This is the issue studied by Ross, Winterhalder, and McElreath (2018) and shown to be highly relevant in Shiner

and Thornbury (2019). Another issue is equally important, however: Ethnic bias is not a binary state. Police officers will vary in ‘how biased’ they are (Holroyd and Sweetman, 2016). Any analysis not taking such variations into account runs the risk of not identifying the true underlying bias dynamics within the police force.

The last issue is a more fundamental one: The above explanations for ethnically biased behaviour are not permissible from a legal standpoint and constitute very narrow definitions of bias. Both outcome and threshold tests quantify a notion of efficiency where searches are sensible as long as the hit rate remains efficient. But this describes a post-search quantity. The law is concerned with whether the officer is justified to initiate the search before knowing what they might find. For example, police officers in the United Kingdom may only initiate a search with a person-specific, concrete suspicion that this individual carries contraband (Home Office, 2014). Within this framework, the hit rate after the search is irrelevant. So while the notion of statistical discrimination appeals to a notion of humans as perfectly rational, fully informed decision-makers considering equilibrium effects, from a legal point of view statistical discrimination is nothing but racial profiling. Similarly, defining ethnic prejudice as an inefficient preference fails to account for the negative cost to society at large in form of the many negative encounters that ethnic minority individuals have with police (Eckhouse, 2018). To address these problems, our work takes a broader definition of bias as well as an officer-specific view.

More precisely, we address these three issues jointly by studying individual police officer decisions. This allows us to avoid pooling information across the widely different police officers that generated the data. As a result, we can go beyond previous work and avoid the assumption of homogeneity across officers and we can model their individual characteristics as they relate to ethnic bias. Lastly, our measure of ethnic bias is broad: We consider over-searching, which we define as the searching of an ethnic group over and above their representation in two baselines of exposure. For each police officer, we infer 1. a patrolling baseline based on where the police officer patrols and 2. a crime suspect baseline based on the crime suspect that a police officer interacts with. The first baseline is the closest approximation of the group most immediately affected by a police officer’s decisions. The second baseline addresses a claim frequently put forth by police: that the ethnic composition of stop and search decisions reflect the ethnic composition of who is committing the crime (e.g., Rudovsky, 2001; Equality and Human Rights Commission, 2013; The Centre for Social Justice, 2018; Bentham,

2021). Benchmarks of criminal involvement have been used in the past (e.g., Gelman, Fagan, and Kiss, 2007), but not at the officer level. For both baselines, if officers were unbiased in their decisions, we would expect their searches to match the baselines.

One may wonder why the literature on officer-specific modelling in the context of ethnic bias is so sparse (for notable exceptions see Ridgeway and MacDonald, 2009; Goncalves and Mello, 2020). The answer is straightforward: even when the data is available, many approaches do not employ an officer-level analysis because individual officers often do not perform many searches in a given year. This leads to small sample sizes and unstable estimates. To ameliorate this issue, our analysis is based on a Bayesian hierarchical model.

Rather than taking the number of searches performed by a police officer as given, we instead estimate search shares—the share of one ethnic group in the officer’s searches. This procedure has the benefit that even if search counts are not sufficiently high, the Bayesian model produces more stable inference. It does so using two levers: First, by virtue of being a hierarchical model, officers in the same team borrow ‘statistical strength’ from each other which leads to overall more stable estimates. Second, we propagate any uncertainty forward. For example, say we have a police officer with very few searches where we cannot reliably characterize the ethnic composition of their searches. The uncertainty around the ethnic composition is then propagated forward into our measure of over-searching for this police officer.

In summary, in spite of having received significant attention, the way that ethnic bias in policing decisions has been modelled in the literature to date has often been insufficiently fine-grained to draw meaningful conclusions. We address this issue by proposing an officer-specific hierarchical Bayesian model that shares information between officers without needing to treat them as a homogeneous mass representing the police force at large.

Modelling domestic abuse data

Domestic abuse is an enormous social issue. In the United Kingdom, an estimated 5.5% of adults over 16 experienced domestic abuse in 2019 (Office for National Statistics, 2019a). It is a complex issue, characterized by a multitude of abusive behaviours ranging from psychological, emotional, financial, physical to sexual abuse.

The social costs of domestic abuse are hard to understate. Victims of domestic abuse suffer physically, mentally and emotionally from the abuse and often incur injuries requiring medical attention (Tjaden and Thoennes, 2000a). In most cases, the abuse is on-going for years (Tjaden and Thoennes, 2000a). The harms of domestic abuse extend to the entire family environment. Children of women suffering from domestic abuse have significantly lower birth weight and often develop psychological and developmental problems (Carlson, 2000; Aizer, 2011).

The incidence of domestic abuse is the result of a complex interplay of structural, family-specific and of spontaneous factors (Jewkes, 2002). For instance, Gracia et al. (2015) show that structural factors such as public disorder, crime levels and residential stability influence the incidence of domestic abuse. Additionally, Freisthler and Weiss (2008) and Livingston (2011) show that the availability of alcohol outlets at the neighbourhood level increase the incidence of domestic abuse. Markham, Doran, and Young (2016) demonstrate a similar relationship to electronic gambling density.

Family-specific factors like the family composition inside the home can strongly influence the incidence of domestic abuse. Tur-Prats (2019) shows that inter-generational cohabitation can decrease domestic abuse: households in which other women, typically the mother of the husband, live with the married couple have lower rates of domestic abuse. Multiple studies have shown that family-specific dynamics such as who holds the economic bargaining power within the household greatly influence the incidence of violence. Aizer (2010) finds that decreasing the gap between male and female wages is associated with a decrease of violence against women. Bobonis, González-Brenes, and Castro (2013) provide even stronger evidence of the importance of economic bargaining power: They find that a cash transfer programme where funds were directly transferred to women reduce the incidence of physical domestic abuse. At the same time, these women are more likely to receive what the authors call “violent threats with no associated abuse” where threats of violence, rather than direct violence, are used to extract money from spouses.

Lastly, domestic abuse is greatly affected by immediate, spontaneous cues. The relationship between major sporting events and domestic abuse is well-documented (Card and Dahl, 2011; Kirby, Francis, and O’Flaherty, 2014; Trendl, Stewart, and Mullet, 2021): Sports tournaments create environments in which different emotions such as sudden excitement, anger or frustration emerge. Additionally, alcohol consumption often plays a prominent role in such tournaments,

resulting in lowered inhibitions which can increase domestic abuse (Wilson, Graham, and Taft, 2017; Leonard and Quigley, 2017; Trendl, Stewart, and Mullet, 2021).

Some work has also focused on identifying factors which would not plausibly account for changes in the incidence of domestic abuse but in the reporting behaviour. Miller and Segal (2019) and Kavanaugh, Sviatschi, and Trako (2019) find that the number of female officers in a police force and the criminal justice system more generally can increase reporting. Muchow and Amuedo-Dorantes (2020) find that heightened awareness of immigration enforcement reduces reports of domestic abuse to police in Los Angeles. Their study provides a partial explanation for the considerable gaps between reporting rates of domestic abuse between ethnic groups.

As the above research shows, domestic abuse is a highly complicated social phenomenon that requires a careful consideration of relevant factors: An investigator needs to think very carefully about the precise data they have available as well as the data-generating mechanism(s). One aspect of this careful thinking again involves data pooling. In fact, the majority of the studies presented so far are based on pooled data.

The typical approach is to bin and count data in geographic units such as census areas and then investigate causal links or correlations to other features. This approach has two problems: First, the geographic area is usually arbitrary. Few social phenomena strictly adhere to administrative unit borders. This leads to a problem called the modifiable area unit problem where the same underlying phenomenon can be interpreted in various ways, depending on the arbitrary drawing of geographical borders (Gehlke and Biehl, 1934). Even more seriously, arbitrary borders induce spatial autocorrelation which significantly affects inference if not accounted for properly (Anselin, 1988; Cordy and Griffith, 1993).

A further issue with pooling data concerns confounding. Recent work has shown that crime is not only affected by the features and demographic composition of the neighbourhood (e.g., Gracia et al., 2015), but that also the reporting of crime is affected by neighbourhood characteristics (Goudriaan, Wittebrood, and Nieuwbeerta, 2006). This means that any relationship estimated between the incidence of reported domestic abuse and neighbourhood features is potentially confounded.

These issues together illustrate that approaches that do not take the process by which a phenomenon is recorded as a data record seriously will be flawed. Indeed, as we will discuss in more detail in Chapter 2, the decision to report a crime is a

complex process shaped by many factors. Clearly, the process by which reports of a crime are generated is different from the process by which crime occurs. This means that researchers cannot naively treat recorded crime as an “imperfect but useful” approximation of the true incidence. While this is true to some extent for all crimes, it is less of a concern with, for example, burglaries where most victims need a crime report to file insurance claims. (Conversely, low-level theft is often under-reported.) For domestic abuse however, the divergence between the process that generated the crime and the process that generated the report of the crime is substantial. While exact estimates are hard to come by, probably only a quarter of victims of domestic abuse report to police (Osborne, Lau, and Britton, 2012). This presents researchers with a challenge: Which data best represents the true prevalence of domestic abuse? Demographic information from police data, domestic abuse support providers and medical institutions show that they all serve different populations (Coy and Kelly, 2011). That suggests that no single institution reaches the full set of people suffering from domestic abuse. In turn, this implies that no one data set can be considered a good approximation of the incidence of domestic abuse. In fact, this is true for many data sets on crime and policing. As we have already seen in Chapter 1, the police officer behaviour itself is a major factor in how stop and search is recorded. Similar things are true for domestic abuse where the behaviour of police officers is a major factor in the decision to report to police, as we discuss in Section 3.3 of Chapter 2. Together, this means that researchers need to take the process by which the data is recorded very seriously when deciding on their statistical approach.

The work presented in Chapter 2 seeks to address the issues presented so far jointly. First, the focus is firmly on modelling the reporting of domestic abuse. This more closely reflects the data actually available.

Second, we use a spatio-temporal point process. This means modelling the actual dynamics of the phenomenon, rather than a discretised version. To try and capture these dynamics, a sophisticated methodological toolkit is required. For starters, one now has to use a continuously valued point process model to avoid pooling of data. Rather than using an implicitly piece-wise constant model in the way a geographic regression would, we instead model the events as the realizations of a non-stationary stochastic rate process. The spatial component of the point process means that we model the spatial location of each report of domestic abuse directly, rather than the rate of reports in some arbitrary unit. We model the timestamp of a report in the same way to account for the fact that crime and reporting vary throughout the day, based on daily routines, daylight,

and season.

An added benefit of this model class is that it allows for differentiating between spatio-temporal clusters and contagion by using a Hawkes-type point process specification (Hawkes, 1971; Ogata, 1988). The conditional intensity of this process depends on two components, a so-called endemic and an epidemic component. The endemic component captures the spatio-temporal nature of domestic abuse reporting while the epidemic component captures the spread of reporting from one report.

With this new modelling approach, we explore the dynamics of crime reporting decisions as expressed in the epidemic component. Many crime types exhibit epidemic behaviour, without much care being taken to distinguish between the crime and its reporting (Loeffler and Flaxman, 2018, is a notable exception). We investigate if reporting is “contagious”, i.e., if the reporting of domestic abuse leads to further reports within the same neighbourhood. Reporting domestic abuse to the police can be quite visible with a police car parked out front and officers inside the home. Given that individuals living in the same neighbourhood tend to be quite similar, we speculate that such visible incidents could lead other victims of domestic abuse to report to police. Whereas domestic abuse behaviours are non-contagious, reporting decisions may well be.

Our hypothesis of localised contagion of reporting was not confirmed by the data. Even though it was plausible to construe the data-generating process of domestic abuse reporting to have a contagion component, we find strong evidence against this hypothesis after evaluating the model. This was only possible because we specified a continuous stochastic model extended from existing approaches that allowed us to explicitly test this hypothesis. This illustrates the significant benefits in a) carefully distinguishing between crime and its reporting and b) in using custom-tailored models for complex hypotheses.

Robust inference under model misspecification

As the first two chapters of this thesis vividly illustrate, statistical modelling in the social sciences is generally a daunting task. This is especially true for modelling crime and illustrated by an example in Chapter 3: There, we are interested in inferring the daily incidence of a type of sexual offence. But reports of sexual violence show extreme spikes on specific dates such as the first of a month which most likely do not reflect spikes in incidence. These outliers are possibly

due to a complex data-generating process about which we can only speculate. For example, media coverage of a high-profile case might lead to a spike in the reporting of cases which have occurred earlier, often years earlier. Without more information, we cannot simply use standard inference procedures without our inferred quantities being heavily influenced by these spikes.

In Chapter 3, we propose a generic robust inference procedure which provides a remedy in such circumstances of model misspecification relative to an unknown data-generating mechanism. In many contexts, it is exceedingly difficult to come up with an approximate description of the data-generating mechanism. For example, practitioners using linear regression do not typically believe that a linear combination of covariates literally generated the outcome of interest. More generally, any statistical model within the social sciences is at best a coarse description of the real world.

Still, we use maximum likelihood estimation or Bayesian inference to infer the parameters of our models. While such likelihood-based statistical procedures have many interpretations, a particularly compelling one is that they minimise a measure of discrepancy between the fitted parametric model and the data-generating mechanism that produced the observations. In fact, the semi-metric minimised by these operations is a statistical divergence first introduced in the seminal work of Kullback and Leibler (1951). In other words, when we optimize the values of our model parameters we minimise the Kullback-Leibler Divergence (KLD) between the model and the data-generating process. While this is the statistically most efficient way of measuring information in a sample for a correctly specified model, under model misspecification this is no longer true. By correct model specification, one commonly refers to a model class rich enough such that there exists a parameter constellation which recovers the data-generating process. Unfortunately, in most of applied research it is rarely possible to identify a statistical model that is capable of perfectly describing the data-generating process in this way. This makes using likelihood-based procedures involving the KLD problematic because one well-known feature of the KLD is that it punishes parameter values that do not fit untypical parts of the data well. While this behaviour is desired in many instances, with model misspecification the robustness of model parameters can be a serious concern.

Particularly important examples of misspecification are contamination, outliers and heterogeneity. Intuitively, this form of misspecification says that while the model fits the bulk of the data well, there is a small part of the data that does not conform to the model. While this may sound like a practically harmless form

of misspecification, it really is not: as we show in Chapter 3, even small amounts of data contamination can disproportionately influence inference outcomes. Recent proposals by Jewson, Smith, and Holmes (2018) address these concerns by proposing the use of divergences other than the KLD. This directly addresses the concerns about model misspecification since these divergences produce robust results. We build on this literature by replacing the standard information measure KLD with the Total Variation Distance (TVD). While the TVD is less statistically efficient, it is robust to misspecification. This can be seen intuitively from the definition of the TVD in Equation (3.2) because the TVD seeks to maximize the average probability mass over the sample space. We trade off a slight decrease in efficiency with significantly more robust inference in the presence of outliers or other kinds of misspecification.

Model misspecification is a particular concern when it comes to modelling discrete data, especially counts. The Poisson distribution rarely fits real-world count data well. In particular, the Poisson distribution has a particular key feature: its variance equals its mean. This assumption is almost always violated with real-world data. Usually, data exhibit over-dispersion which means that the variance is higher than the mean. A well-known feature of inference with Poisson models is that ignoring the presence of over-dispersion leads to optimistic standard errors of the estimators (Cameron and Trivedi, 2013). One can correct this by using robust standard errors. Typically however, alternative models with additional parameters that explicitly model over-dispersion such as the Negative Binomial are used. Often, data also exhibit ‘excess’ counts at particular points in the distribution, usually zero. Then, zero-augmented models such as hurdle or zero-inflated models are available to accommodate high frequencies of specific counts with further additional parameters. These models are sufficiently different that they each entail very different distributions, interpretations and outputs (Cameron and Trivedi, 2013). Choosing between them is difficult and highly domain- and context-dependent (see discussions in Lindsey and Jones, 1998; Gao and Khoshgoftaar, 2007; Ver Hoef and Boveng, 2007; Sileshi, Hailu, and Nyadzi, 2009; Casals, Girabent-Farres, and Carrasco, 2014; Hawinkel et al., 2020).

In some applications, the use of these models helps account for specific features of the data-generating process (e.g., Lambert, 1992). Often however, the overdispersion of observed data is simply the result of randomness. Then, the use of such misspecified models can seriously threaten the robustness of our inferred parameters.

Beyond these subtle features strongly influencing model selection and subsequent analysis, count model inference shares some problems with more general inference problems. As with most other settings, outliers in higher dimensions are difficult to detect, leading to a model misspecification problem: The candidate model f is misspecified for those observations.

Without further knowledge about the type of misspecification, a generalised robust inference approach is desirable. Using robust divergences such as the TVD is a black-box approach to this very specific inference problem. With no additional information, this approach robustifies analyses and results to model misspecification. Robustifying the inference of discrete data models in particular is appealing for two reasons: 1. It is an abundant yet challenging problem in the applied sciences and 2. the discrete nature of the outcome variable allows for an elegant solution and exciting theory development.

In Chapter 3 we apply this problem to reports of serious sexual offences. The example is highly stylized and instructive to show the kind of applied problem where our method is most useful. In the example, we are interested in inferring the daily incidence of a type of sexual offence in the presence of extreme outliers. Most likely, these outliers are the result of back-dating historical reports but this is speculation. Thus, a generically robust method is useful here.

The example is also an illustration of the larger theme of this thesis: Whenever we use data sets on human behaviour in the field, we have to think very carefully about the information contained in the data. Is the behaviour of interest captured without distortion or selection? Are existing approaches modelling the phenomenon at the resolution at which it is occurring? Such questions are particularly important when it comes to modelling police data where we as scientists are trying to understand complex phenomena and potentially propose solutions to important social issues. As the diversity of methodology in this thesis shows, there is no universal answers, only context- and domain-specific solutions.

Chapter 1

Officer bias, over-patrolling, and ethnic disparities in stop and search

1.1 Introduction

Ethnic minorities are over-represented in police searches compared to White people. In England, Black and Asian people make up 11% of the population, yet they account for 30% of all English police searches, called stop and search (Home Office, 2018a).

Search decisions come with considerable consequences: searches can create feedback loops where an individual is repeatedly searched because they were searched in the past (Quinton, 2011; Brayne, 2017) which increases their likelihood of being arrested, thereby creating further feedback loops in the criminal justice system (May, Gyateng, and Hough, 2010; Kohler-Hausmann, 2013). High levels of searches further result in diminished citizen engagement with police, diminished political engagement, reduced perceptions of police legitimacy and trust in police (Sharp and Atherton, 2007; Lerman and Weaver, 2014; Tyler, Fagan, and Geller, 2014; Bradford, 2015; Bradford, 2017; Laniyonu, 2018). In addition, invasive search encounters can result in psychological harm to searched individuals, leading to increased symptoms of stress, anxiety and trauma (Skogan, 2006; Geller et al., 2014; Delsol, 2015).

It is therefore crucial to understand the reasons for the over-searching of ethnic minorities. Here we explore ethnic bias in search decisions at the officer level by focusing on individual officers' bias and the factors shaping these biases.

Our approach is two-fold. First, we investigate officers' search biases against an ethnic group relative to two officer-specific baselines: the ethnic composition of crime suspects and of the areas they patrol. Second, we then examine the contributions of officers' search biases and of biases in deployment decisions to the over-representation of ethnic minorities in stop and search.

We demonstrate that the majority of officers over-search Asian and Black people, whichever baseline we compare their searches against. Our results show that officers perform more searches of ethnic minorities than can be explained by the ethnic composition of the areas officers patrol or of the crime suspects officers interact with. However, over-searching by individual officers cannot account for all of the over-representation of ethnic minorities in stop and search. Over-patrolling is part of it: The median officer in our sample patrols areas which are 1.16 times more Asian and 1.37 times more Black than the West Midlands police force area. In other words, police officers are deployed to more ethnically diverse areas. Such deployment decisions contribute to the over-searching of ethnic minorities (Sampson and Lauritsen, 1997; Shiner, Carre, et al., 2018). We find that these biases in deployment decisions multiply with individual officers' biases. Both together account for the overall bias against ethnic minorities in stop and search.

Such over-searching or bias is not equivalent to discrimination. Conclusively attributing empirical patterns of disparities to ethnic or racial discrimination is challenging (Neil and Winship, 2018; Simoiu, Corbett-Davies, and Goel, 2017; Knox, Lowe, and Mummolo, 2020). We believe it is nonetheless important to uncover, document and dissect ethnic disparities because differential rates of contact with police entail far-reaching consequences for the criminalisation of ethnic minority groups and, not least, the legitimacy of the institution of police. In our study we make two important contributions to the literature on ethnic bias in policing: First, we provide officer-specific measures of search bias relative to the crimes suspects an officer encounters and relative to the population in the area the officer patrols. Second, we find that officers' search biases are smaller than search bias on the police force level, suggesting that deployment decisions contribute to the overall search bias against ethnic minorities in stop and search.

1.2 Related work

Our approach and results connect to a rich literature on stop and search, and on ethnic bias in policing more generally. Stop and search in the United Kingdom is a widely used policing power characterised by police forces as a crucial tool to prevent and investigate crime (Shiner, 2010; Equality and Human Rights Commission, 2013; The Centre for Social Justice, 2018). If achieving these aims justifies persistent ethnic disparities has been powerfully challenged in the landmark Scarman and MacPherson reports: They rejected police explanations for disproportionate use of stop and search and instead described it as a prime example of institutional racism (Scarman, 1981; Macpherson, 1999; Delsol and Shiner, 2006; Bowling and Phillips, 2007). Furthermore, the empirical evidence suggests that stop and search has, at best, only minor effects on crime. Most studies, especially those conducted in the United Kingdom, do not find any evidence of crime reductions in response to stop and search (Ward, Nicholas, and Willoughby, 2011; Delsol, 2015; McCandless et al., 2016; Weisburd et al., 2016; MacDonald, Fagan, and Geller, 2016; Tiratelli, Quinton, and Bradford, 2018).

Police frequently attribute the over-representation of ethnic minorities in stop and search to their over-representation in crime, implying that ethnic minorities perpetrate more crime than White people (Equality and Human Rights Commission, 2013; The Centre for Social Justice, 2018; Quinton, Bland, and Miller, 2000; Rudovsky, 2001; Phillips and Bowling, 2007; Shiner, 2010; Her Majesty’s Inspectorate of Constabulary, 2013). This argument can result in a self-fulfilling prophecy because the process of observing and recording crime already depends on the wider social context of policing. In this context, deployment decisions (Sampson and Lauritsen, 1997; Smith, 1986; Elliott et al., 1995; Fagan and Davies, 2000; Kane, 2003), arrest probabilities (Kohler-Hausmann, 2013; Lammy, 2017) and the accurate recording of crime (Gounev and Bezlov, 2006; Richardson, Schultz, and Crawford, 2019) are not independent of ethnic group. As a consequence, crime data are not an objective benchmark of true criminal behaviour. In our work we do not take this into account for a simple reason: We are interested whether police officers’ actions match the benchmarks they assemble themselves. Thus we compare officers’ stop and search decisions to their own encounters with crime suspects.

In light of police’s potential to criminalise minorities, the role of ethnic bias in police decision-making deserves further inquiry. Police officers operate within the tension between their roles as individual decision-makers and agents of the

institution of the police, influenced by the organisational protocols and structures (Mawby and Wright, 2008; Shiner, 2010; Oberfield, 2012). At the individual level, there is ample evidence of biased attitudes held by police officers as well as of racially or ethnically motivated behaviour (Smith and Gray, 1985; Waddington, 1999; Eberhardt et al., 2004; Alpert, MacDonald, and Dunham, 2005; Warren et al., 2006; Correll et al., 2007; Morris, Burden, and Weekes, 2004; American Civil Liberties Union, 2009; Adebawale, 2013; Quinton, 2015). Ethnic bias at the institutional level is equally important. The Stephen Lawrence Inquiry in 1999 with its emphasis on institutional racism has sparked a varied discussion on the role of police forces in creating ethnic disparities in the United Kingdom (Lea, 2000; Reiner, 2010; Shiner, 2015). Two factors have been highlighted in particular: First, structures within the police force perpetuate and broadcast biased beliefs through various hierarchies (Lea, 2000; Bowling and Phillips, 2007; Shiner, 2010; Shiner, 2015). Second, deployment decisions by the police force—that is, decisions about which areas to prioritise and deploy officers to—are under scrutiny, given that these decisions can create disparities at the population level, independently of how individual officers behave (Delsol and Shiner, 2006; Bowling and Phillips, 2007). Deployments are also often targeted at specific behaviours such as drug use in specific neighbourhoods, often deprived and ethnically diverse. These types of targets raise concerns about the criminalisation of minority communities (Delsol and Shiner, 2006; Shiner, Carre, et al., 2018).

Officer teams are intermediaries between officers and the police force, often with their own norms and cultures (Mawby and Wright, 2008; Reiner, 2010). A recent study noted remarkable differences between different teams within the same English police force: Teams tasked with proactive policing not only performed the highest number of searches within the force but were also over-searching Black people at higher rates than other teams (Shiner and Thornbury, 2019). In our analysis we explore the relevance of officer teams by accounting for differences between teams and by including the ethnic composition of officers’ teams into our model.

The tension between individual and institutional behaviour also applies to other policing activities such as drug enforcement (Kohler-Hausmann, 2013; Shiner, Carre, et al., 2018), arrests (Sekhon, 2018) and use of force (Ross, 2015). The literature on use of force in particular is currently debating an important consequence of this tension: What is the appropriate level of analysis of use of force data? We will briefly outline this debate since the analysis of stop and search data is characterised by the same tension and because our results can directly speak

to an ongoing discussion within the use of force literature. In the United States, Black people are subject to higher rates of police use of force, particularly lethal use of force, than White people relative to their shares in the population (Ross, 2015; Edwards, Esposito, and Lee, 2018). Some studies have argued that the general population in an area is not the appropriate comparison: Instead one should compare rates of use of force to how often Black and White people come into contact with police (Fryer, 2019; Cesario, Johnson, and Terrill, 2019; Johnson, Tress, et al., 2019). After conditioning on the rate with which police encounter Black individuals, Fryer (2019) finds a reversal of ethnic disparities: Police are apparently less likely to employ lethal force on Black people than White people.

An issue with this approach is pooling: Fryer (2019)’s analyses are at the police department level, pooling all officers together. However, if officers are not homogeneous and differ in how often they encounter Black people or differ in how biased they are against Black people, then pooling their data can lead to erroneous conclusions. This phenomenon, called Simpson’s paradox, is explicitly considered by Ross, Winterhalder, and McElreath (2018): In response to Fryer (2019), they develop a generative model where all officers are biased against Black people but differ in how often they encounter them. Already a small group of officers which encounters Black people at high rates is sufficient to confound the pooled analysis and point toward anti-White biases (when in fact all officers exhibit anti-Black bias by construction). In other words, pooled analyses of use of force data can fail to detect ethnic bias with heterogeneous police officers (Ross, Winterhalder, and McElreath, 2018; Simpson, 1951; Neil and Winship, 2018).

The pooling problem directly applies to pooled analyses of police searches. If officers differ in how often they encounter criminals of different ethnicities, then a police department-level analysis of searches conditioned on crime can be confounded and fail to identify the direction of the disparity. Generally, analyses of searches tend to find over-representation of ethnic minorities even after conditioning on crime. For example, Gelman, Fagan, and Kiss (2007) find over-representation of Black and Hispanic people in pedestrian stops-and-frisks in New York City after adjusting for race-specific representations in crime, a pattern substantiated in other analyses (Fagan and Davies, 2000; Ridgeway, 2007). In addition to pedestrian searches, traffic stops—where similar ethnic biases persist (Pierson et al., 2020)—are often compared to benchmarks of criminal behaviour (Alpert, Smith, and Dunham, 2004; Rojek, Rosenfeld, and Decker, 2012; Withrow and Williams, 2015). All of these analyses are performed at the police department level meaning that they could be potentially confounded.

Internal bench-marking is an officer-specific approach which matches each police officer to similarly-situated officers (Ridgeway and MacDonald, 2009). The officer’s behaviour is possibly problematic if it deviates substantially from their peers’. A drawback of this method is that it can only reveal individual officers’ biases relative to their peers. For example, only 15 out of 2,756 officers of the New York City Police Department are flagged as potentially biased (Ridgeway and MacDonald, 2009), far too few to explain the overall level of over-representation of ethnic minorities in stop and search.

In our study, we explore stop and search behaviour at the level of the individual officer, following a panel of officers over time. We compare an officer’s searches of an ethnic group to the officer’s direct experiences of the crime involvement of this group. By not pooling our data we thereby circumvent the issue of Simpson’s paradox.

1.3 Data

Our data consists of records of searches between 01/04/2014 and 30/09/2018 provided by West Midlands Police in England as well as all recorded crimes in the same period. We split this time frame into nine periods of 6 months each, beginning from 01/04/2014. We chose this time resolution because periods shorter than 6 months result in sparse officer-level information. Officers which performed searches in fewer than 50% of the half-year periods, i.e., in fewer than five half-year periods out of the nine in our study period, were excluded to avoid data sparsity issues. The final file covers 1,194 officers observed in 29 teams, 203,176 reported crimes and 36,028 searches.

Our analysis is focused on so-called suspicion searches. In the United Kingdom, police officers routinely stop and question members of the public. During these unrecorded conversations, officers can ask individuals to account for themselves. If at any point the officers form a ‘reasonable suspicion’ that the person is in possession of illegal items such as weapons, drugs or burglary tools or in possession of stolen items, officers can initiate a search of the person’s clothing and belongings (Home Office, 2014).

At this point, the encounter must be recorded in the form of a stop and search record detailing information about the searched person and the officer’s justification for the search. At the end of the search encounter, the searched person has to be supplied with a reference number to the record. A search may

be initiated only under powers requiring ‘reasonable grounds for suspicion’ as detailed or with prior authorisation. In our analysis we restrict our attention to suspicion searches, which account for 99.4% of all searches, because only these searches are initiated at the discretion of the searching officer.

In our analysis we use self-defined ethnicity, which is someone’s response to the question “What is your ethnic group?”. We focus our analysis on Asian, Black and White people because sample sizes are too small for the remaining Mixed, Chinese or Other groups.

In our analysis we rely on two officer-specific baselines: the crime suspects an officer encounters and the residents in the officer’s patrolling area. We obtain the crime suspect information by linking officers to the reported crime cases they responded to and then counting the person(s) suspected by police of having committed the offense. For the patrolling information, we calculate how often an officer visits a given geographical census unit using additional patrolling data and obtain an officer-specific patrol intensity share for the area. We use the smallest geographical unit provided by the 2011 Office for National Statistics (ONS) census, 2011 Output Areas (Office for National Statistics, 2016). We then multiply the number of residents in each census unit with the intensity share and sum them to obtain patrolling intensity-weighted counts of the residents in an officer’s patrolling area. Our data form a panel of search counts, crime suspect counts and patrol counts for each officer over 9 half-year (6 month) intervals.

Altogether, we use the following variables: counts of officers’ searches; counts of officers’ crime suspect encounters; counts of residents in officers’ patrolling areas, all broken down by ethnic group; officer gender (dummy encoded); officer age; officer experience and two dummy variables indicating whether officer i is Asian or Black. We standardise officer age and officer experience to have mean 0 and standard deviation 1. We summarise these variables in Table 1.1.

Officers transfer between teams during our study period. We account for this in our model with the team-specific intercept α_j . All officers are assigned to the team j they were part of for the majority of the time in each 6-month period.

1.4 Methods

We define two measures of police officer over-searching which we define in Section 1.4.1. To obtain the shares on which our two measures are based, we use

Variable	Mean	SD	Min	Max
Time-varying variables per half-year				
Search counts				
Asian	2.41	4.55	0.00	76.00
Black	1.72	4.13	0.00	94.00
White	5.39	9.18	0.00	125.00
Suspect counts				
Asian	10.61	12.30	0.00	80.00
Black	8.29	8.87	0.00	97.00
White	43.41	40.71	0.00	212.00
Patrolling counts				
Asian	85.56	46.95	0.00	349.00
Black	32.30	19.45	1.00	145.00
White	200.84	49.62	37.00	343.00
Officer age in years	37.78	7.34	19.08	62.00
Standardized officer age	0.00	1.00	-2.55	3.30
Officer experience	10.81	5.25	0.17	30.42
Standardized officer experience	0.00	1.00	-2.02	3.73
Fixed variables				
Female officer	0.12	0.32	0.00	1.00
Asian officer	0.06	0.25	0.00	1.00
Black officer	0.01	0.10	0.00	1.00
Number of observed half-years per officer	8.46	1.14	5.00 ¹	9.00
Share of White officers in team	0.93	0.06	0.81	1.00
Total number of observations (officers \times half-years)	N =	10,103		
Number of officers	N =	1,194		
Number of teams	N =	29		

¹ We exclude officers with fewer than 5 half-years' worth of observations (see Data section)

Table 1.1: Means, standard deviations (SD), minima and maxima of variables used in the multinomial model in Section 1.4.2.

a hierarchical Bayesian multinomial regression model which we describe in Section 1.4.2.

1.4.1 Measures of over-searching

We infer search shares p , crime suspect shares ζ and patrol population shares ρ for each officer i in time period t from the respective counts using the model described in the next section. These shares represent the share of each ethnic group e in the officer's searches, crime suspect encounters and patrol counts, respectively.

They form the basis for our two measures of over-searching:

1. O^S , officer over-searching relative to crime suspects. For each officer we obtain O_{ite}^S by dividing the officer's search share p of ethnic group e in time period t by the officer's suspect share ζ of e in t . If O^S is larger than 1 then the officer over-searches an ethnicity relative to how often they encounter the ethnic group as crime suspects. If O^S is smaller than 1 then the officer under-searches an ethnic group relative to suspects and if O^S is exactly 1 then the officer searches that ethnicity at the same rate as they appear in the officer's crime suspects.
2. O^P , officer over-searching relative to patrol. For each officer we obtain O_{ite}^P by dividing the officer's search share p of ethnic group e in time period t by the officer's patrol share ρ of e in t . O^P has the same interpretation as O^S : If O^P is larger (smaller) than 1 then the officer over-searches (under-searches) that ethnic group relative to the ethnic composition of the area they patrol.

For example, for the median officer Asian people make up 23% of their searches, 15% of the crime suspects they interact with and 23% of the areas the officer patrols. Officer over-searching of Asian people relative to crime for this officer is $O^S = 0.23/0.15 \approx 1.53$ which means that the officer over-searches Asian people relative to crime suspects by a factor of 1.53. Officer over-searching relative to patrol for this officer is $O^P = 0.23/0.23 = 1$, meaning that this officer searches Asian people about as much as they encounter Asian people on patrol.

1.4.2 Multinomial model

Every officer in our sample performs a number of Asian, Black and White searches in a given time frame. We are then interested in characterising the composition

of the searches by an officer: Which percentage of the officer’s searches were searches of Asian people? To do this, we employ a Bayesian Multinomial logit model (sometimes also called a Softmax model) where the search shares are a non-linear combination of officer characteristics such as age and team characteristics such as team composition. We then repeat this procedure to characterise the ethnic composition of the officer’s patrolling area and interactions with crime suspects. Based on these three officer- and time frame-specific shares (searches, patrol and crime) we then build our two measures of disparities of relative to crime suspects and patrol.

More formally, our data are counts of searches of ethnic group e by officer i in time period t . For each officer we thus have a vector $Y_{it} \in \mathbb{N}_0^E$ where $E = 3$ are the three ethnic groups we consider: Asian, Black and White, which we abbreviate to A, B, W for ease of notation.

We are then interested in the proportions of each ethnic group in the total number of searches by officer i in t as a function of covariates. Formally, we model the allocation of the total number of searches by i in t , $\sum_{e \in \{A, B, W\}} Y_{ite}$ (shortened to $\sum_e Y_{ite}$ for ease of notation), into $E = 3$ ethnic groups as follows:

$$Y_{it} \sim \text{Multinomial} \left(\sum_e Y_{ite}, p \right), \quad p = \text{Softmax}(\theta_{it}). \quad (1.1)$$

In words, Equation (1.1) states that each observation vector Y_{it} is modelled by the vector $\theta_{it} \in \mathbb{R}^E$ where θ_{it} gives an officer’s propensity to search ethnic group e as a function of some covariates. To obtain valid proportions, we use the $\text{Softmax}(\cdot)$ function which normalises a vector of real numbers into a vector of proportions that sum to 1. This means that $p = \text{Softmax}(\theta_{it})$ gives the share of each ethnic group e in $\sum_e Y_{ite}$, the quantity of interest.

However, θ_{it} is not yet identifiable because the same values of $p = \text{Softmax}(\theta_{it})$ can be induced by different θ_{it} . This is easily resolved by setting $\theta_{it \text{ White}} = 0$. In doing so, $\theta_{it \text{ Asian}}$ and $\theta_{it \text{ Black}}$ then represent an officer’s propensity to search Asian or Black individuals relative to searching White people and θ_{it} is uniquely identified.

We model θ_{it} as a function of the demographic covariates listed in Table 1.1. The coefficients of these covariates represent their relative contribution to an officer’s propensity to search Asian or Black people over White people. In modelling θ_{it} we are particularly interested in the contribution of an ethnic group’s proportion in the officer’s crime suspect population and the contribution of an ethnic

group’s proportion in the officer’s residential population in the patrolling area.

We observe a vector of counts of crime suspects and a vector of counts of residents encountered on patrol. We then infer the proportions of each group in those vectors. To this end, we introduce four additional terms: S_{it} , ζ_{it} , P_{it} and ρ_{it} . Similarly to Y_{it} , $S_{it} \in \mathbb{N}_0^E$ is a vector holding counts of crime suspect encounters by officer i in t for $E = 3$ ethnic groups. Because we do not use any covariates to model the allocation of S_{it} , we can directly model the proportions rather than using the $\text{Softmax}(\cdot)$ transformation from before. ζ_{it} is the vector directly giving the suspect shares, that is the proportions of each ethnic group e in S_{it} . The remaining two terms follow the same logic: P_{it} gives counts of residents encountered on patrol by officer i in time period t . ρ_{it} directly models the patrol shares—the proportions of each ethnic group in P_{it} . More formally,

$$\begin{aligned} S_{it} &\sim \text{Multinomial} \left(\sum_e S_{ite}, \zeta_{it} \right), \\ P_{it} &\sim \text{Multinomial} \left(\sum_e P_{ite}, \rho_{it} \right). \end{aligned}$$

Taken together, this corresponds to the following model:

$$\begin{aligned} \theta_{it \text{ Asian}} &= \alpha_{j[it]A} + \beta_A x'_{iteA} + \gamma_A \zeta_{itA} + \delta_A \rho_{itA} + \omega_A w_{j[it]} \\ \theta_{it \text{ Black}} &= \alpha_{j[it]B} + \beta_B x'_{iteB} + \gamma_B \zeta_{itB} + \delta_B \rho_{itB} + \omega_B w_{j[it]} \\ \theta_{it \text{ White}} &= 0, \end{aligned}$$

where $\alpha_{j[it]e}$ is an ethnicity-specific group-level intercept corresponding to the team j that officer i was part of in time period t . x'_{ite} is a vector holding i ’s covariate information at t specific to ethnic group e . $w_{j[it]}$ gives the share of White officers in the team officer i was in in time period t .

Modelling suspect and patrol shares as the allocation of suspect and patrolling counts allows us to account for measurement error. For example, if an officer encounters only few crime suspects, then the uncertainty in the suspect shares will be large because the estimates are based on few data points. The uncertainty in the shares will then be propagated forward to the inference on γ and δ such that noisier, less certain shares receive less weight than shares inferred from sufficient amounts of data.

We specify prior distributions on model parameters as follows: The group-level

intercepts $\alpha_j \sim N(\mu_\alpha, \sigma_\alpha)$ where $\mu_\alpha \sim N(0, 1)$ and $\sigma_\alpha \sim N^+(0, 1)$ (half-normal) and all regression coefficients $\beta, \gamma, \delta, w \sim N(0, 2)$. For ζ we use weakly informative Dirichlet priors parametrised with the respective share of each ethnic group in all arrests in England and Wales in the year 2016/17. (The Home Office does not publish crime by ethnicity.) This yields the prior $\zeta \sim \text{Dirichlet}(0.43, 0.61, 5.00)$ corresponding to country-wide shares of (0.07, 0.10, 0.82). Similarly, for ρ we use the share of each ethnic group in England in the 2011 ONS census: $\rho \sim \text{Dirichlet}(0.39, 0.21, 5.00)$ which corresponds to shares of (0.07, 0.04, 0.89) (Home Office, 2017; Office for National Statistics, 2011a).

We fit the full model with **Stan** in R version 3.6.3 using **rstan** version 2.19.3 (Carpenter et al., 2017; Stan Development Team, 2020). Hamiltonian Monte Carlo sampling was performed on four chains with each 1,000 warm-up draws and 1,000 sampling draws, resulting in 4,000 draws from the posterior distribution in total.

The fit of the model to the observed data is checked in Figure 1.8 in Appendix 1.B. All code used to produce the results is available online at https://github.com/laravomfell/ethnic_bias_stop_and_search. Since the original data from West Midlands Police may not be shared publicly, the repository includes a file `code/generate_synthetic_data.R` which generates synthetic data. The distributions of the variables in the synthetic data match the distributions in our data.

1.5 Results

We perform Bayesian inference. Before seeing the data, we have prior information about likely values of the parameters which are updated with the likelihood of the data to obtain the posterior distribution. A sample, sometimes also called draw, from the posterior is a plausible parameter value consistent with the prior information and observed data. We provide 90% uncertainty intervals for the parameters, sometimes also called credible intervals (Kruschke and Liddell, 2018). 90% of our posterior distribution over the parameter lies within the 90% uncertainty interval.

We present our results in three parts: (i) estimation of search shares, (ii) measures of over-searching O^S and O^P and (iii) the discrepancy between officer-level and force-level search bias.

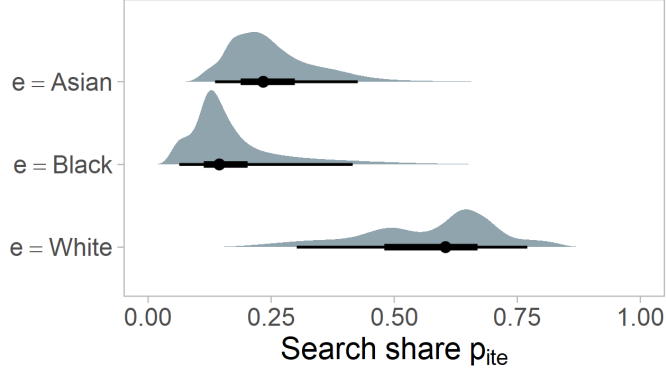


Figure 1.1: Posterior densities of search shares p_{ite} over all 1,194 officers and 9 time periods from the full model, resulting in 10,103 observations. Search shares are the proportion of each ethnic group in the officer’s searches. The black dot represents the medians of the distributions aggregated over e and t and represent search shares for the median officer. Black lines show 50% and 90% uncertainty intervals which represent the spread of behaviour by 50% and 90% of the police officer workforce. There are multiple modes in the posteriors which simply means that there are different clusters of officers with similar search shares. (For a version disaggregated by time see Figure 1.6 in Appendix 1.B.)

1.5.1 Inference of search shares

We infer p_{ite} , the share of each ethnic group e in officer i ’s searches in time period t , as a function of the officer’s suspect shares and patrol shares in time period t , and their gender, age, experience, ethnic group and the share of White officers in their team.

Figure 1.1 shows the posteriors of p_{ite} for each ethnic group over all officers and time periods based on the full model. Due to the aggregation over officer-specific posteriors they represent the (posterior) behaviour of the entire workforce of searching officers and show that searches by the median police officer are 23% Asian, 13% Black and 65% White (with the remainder due to rounding).

As explained above, we infer the search shares as a function of officer and team characteristics and the officer’s suspect and patrol shares. To do this, we first infer each officer’s propensity to search Asian and Black people, called θ_{Asian} and θ_{Black} , and then transform these propensities into search shares. In Figure 1.2, we show the posteriors of these coefficients. We find no credible evidence that officer age and ethnicity are associated with search shares. Officer gender and experience play a minor role where female or experienced officers search fewer ethnic minorities. Relative to the other associations, they are scarcely meaningful.

Instead, we find associations of search shares with officer-level suspect and

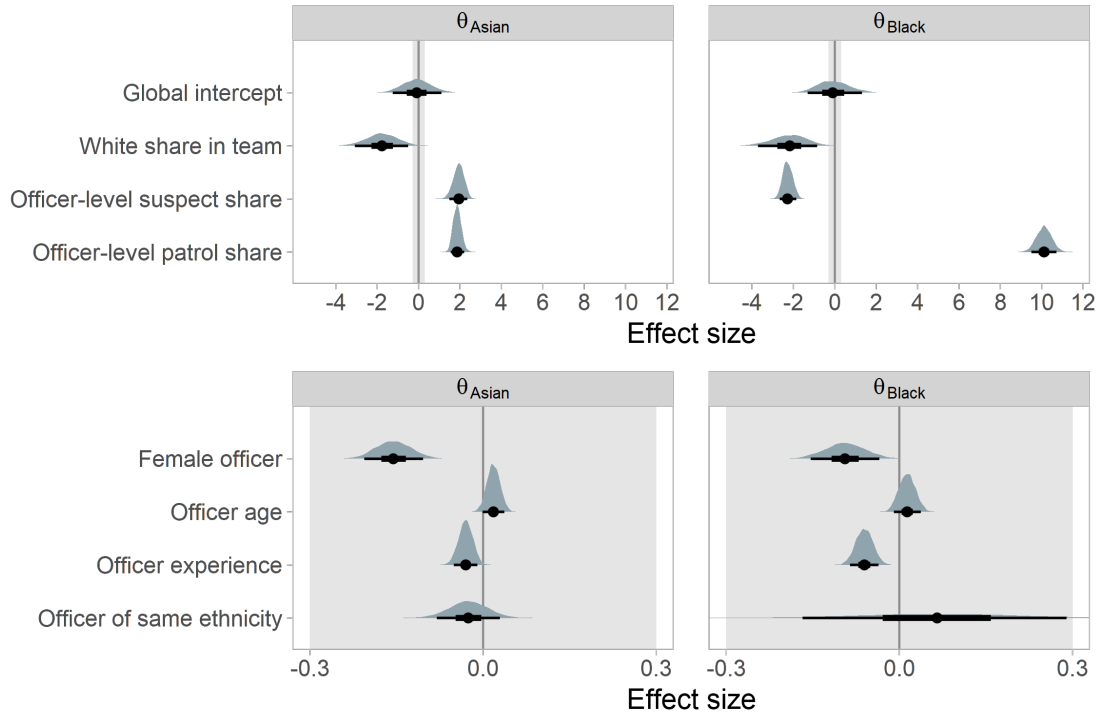


Figure 1.2: Posterior densities of the coefficients used to infer search shares p . A positive effect on θ_{Asian} means a larger Asian search share. Similarly, a positive effect on θ_{Black} implies a larger Black search share. Black dots show the median of the posteriors while black lines show 50% and 90% uncertainty intervals which contain 50% and 90% of the posterior distribution. For visual clarity we show the associations of officer gender, age, experience and ethnicity on a zoomed in scale of $[-0.3, 0.3]$ compared to the others. The figure is based on 10,103 observations. Table 1.3 gives the medians and 90% uncertainty intervals.

patrol shares. The association with Asian suspect shares is positive, meaning that officers with a higher share of Asian crime suspects also have a higher Asian search share. Interestingly, the association with Black suspect shares is negative, meaning that officers who encounter more Black crime suspects have lower Black search shares. The association with patrol shares is more intuitive: the ethnic composition of searches reflects that of the areas officers patrol. In principle, this admits two competing hypotheses: Either, officers are searching at random or they explicitly adjust for the population in their patrol areas. As we demonstrate in the next section however, officers over-search ethnic minorities relative to their patrolling areas which suggests that officers do not search at random.

Last, we comment on our team-level results. In predominantly White teams, Asian and Black people make up a lower share of searches than in more ethnically diverse teams. While it would be preferable to differentiate between Black and Asian officers, we have to treat them as a single group in the analysis as there is insufficient variance in the separate shares of Black and Asian officers in a team, due to the lower numbers of non-White officers in our sample. Our Bayesian model includes team-specific intercepts to account for differences in search shares between teams. The results show that the ethnic composition of searches varies considerably between teams as evidenced by the intercepts' standard deviations. Specifically, they are 0.35 (90% UI [0.27, 0.46]) for Asian searches and 0.43 (90% UI [0.34, 0.57]) for Black searches. Presumably, these differences are due to team specialisation, as officers' routines are determined by their responsibilities.

1.5.2 Measures of officer over-searching

Next, we discuss officer over-searching relative to crime suspects (O^S) and relative to patrol (O^P). We first show draws from the posterior distributions of O^S and O^P in Figure 1.3. Again, the distributions represent the aggregate over officer-specific posteriors and, as such, the behaviour of the entire workforce of searching officers in our sample. The median officer over-searches Asian people by a factor of 1.57 (90% UI [0.80, 6.64]), Black people by a factor of 1.21 (90% UI [0.30, 5.83]) and under-searches White people by a factor of 0.85 (90% UI [0.52, 1.32]) relative to suspects. The uncertainty intervals for Asian and Black searching are wide on the aggregate because they also are wide on the officer level. The interpretation is that we are uncertain about the precise level of officers' search bias against ethnic minorities relative to suspects, but officers are more likely to over- than under-search Asian and Black people. In contrast, the results for

White disparities are clear: More than half of officers under-search White people relative to suspects.

We can be more confident about the actual levels of over-searching relative to patrol. The right-hand side of Figure 1.3 shows that the median officer over-searches Asian people by a factor of 1.01 (90% UI [0.57, 2.18]), Black people by a factor of 1.69 (90% UI [0.946, 3.97]) and under-searches White people by a factor of 0.89 (90% UI [0.64, 1.30]) relative to patrol.

The summaries of O^S and O^P presented so far are coarse: They only allow us to make statements about the aggregate of all officers. To refine the resolution, we compute the posterior probability that an individual officer over-searches a particular ethnic group, both relative to suspects and patrol from the posteriors of the officer-specific disparities. We do this by calculating for each officer how many of the posterior draws of the officer-specific over-searching distributions O^S and O^P from our model are above 1. For example, if this probability is 1, then the officer always over-searches. Similarly, if this probability is 0.5, the officer's search shares perfectly match the suspect or patrol baselines.

Figure 1.4 shows histograms of these probabilities for all officers. The left-hand side is in line with what we have already seen on the aggregate in Figure 1.3: Most officers over-search Asian and Black people while virtually all officers under-search White people relative to suspects. However, the right-hand side of Figure 1.4 reveals a pattern that would be left obscured by only studying the aggregate. Particularly, we observe a split between officers: Some officers under-search Asian people, while others consistently over-search them relative to patrol. Since these officer groups are of roughly the same size, the aggregate incorrectly suggests that officers do not over-search Asian people. In contrast, the officer-level results for Black and White people match the aggregate: Virtually all officers over-search Black people relative to patrol. In fact, 69% of the officers have a posterior probability of over-searching Black people that exceeds 0.95. Similarly, the vast majority of officers under-search White people relative to patrol. There is no change and no discernible dependence in O^S and O^P over time, a point we explore in more detail in the appendix.

1.5.3 Officer- compared to force-level bias

Last, we discuss the implications of our officer-level results on the overall over-representation of ethnic minorities in stop and search. The median officer patrols more ethnically diverse areas than are representative for the police force's area of

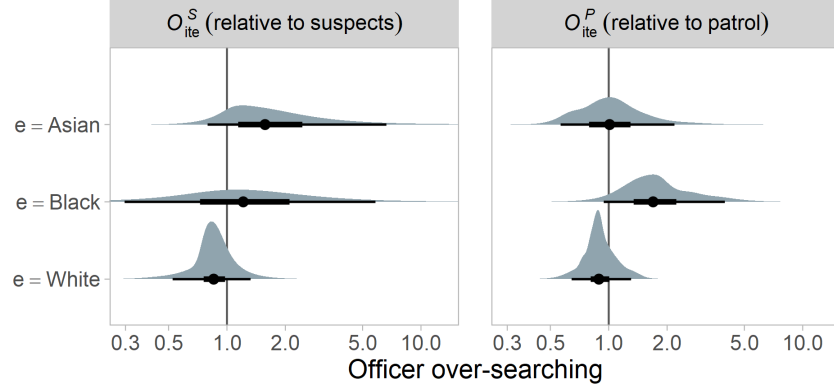


Figure 1.3: Posterior densities of O_{ite}^S and O_{ite}^P aggregated over all officers and time periods. The distributions represent the behaviour of the entire workforce of searching officers. The black dots represent the median officers and the black lines represent 50% and 90% uncertainty intervals. Note that the x-axes of both panels are on the log-scale. For visual clarity, we only show values between $[0.3, 13]$. 1.5% of all posterior probability is excluded by this choice. The figure is based on 10,103 observations. (For a version disaggregated by time see Figure 1.7 in Appendix 1.B.)

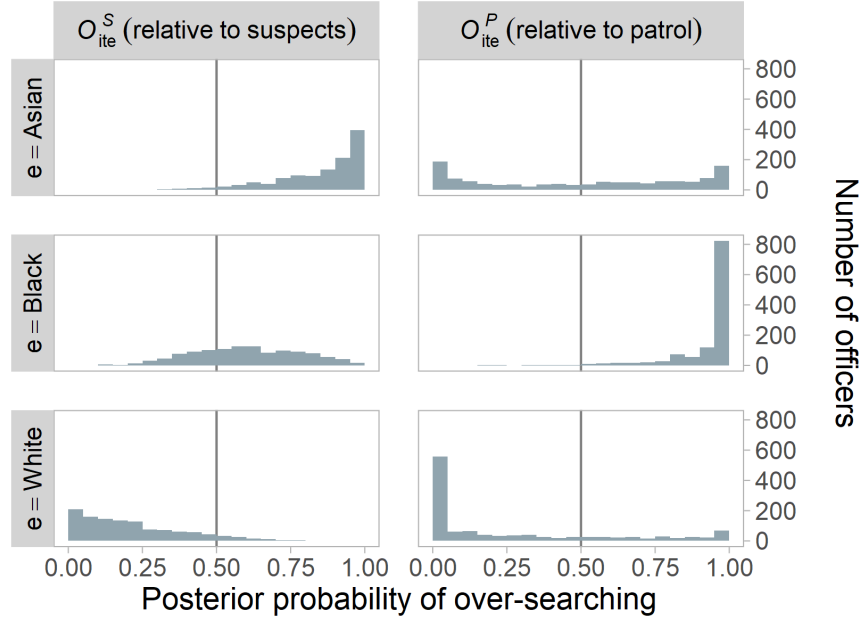


Figure 1.4: Histograms of the posterior probabilities of O_{ite}^S and O_{ite}^P above 1 for each officer where the posterior probability gives how many of the 4,000 posterior draws from an officer-specific distribution are above 1. If the posterior probability above 1 for an officer is 1, the officer always over-searches an ethnic group. If the posterior probability above 1 for an officer is 0.5 then the officer's search shares perfectly match the suspect or patrol baselines. The figure is based on 10,103 observations.

operation. For the remainder of the analysis, we only consider over- and under-searching relative to the patrolling baseline. This is because while police forces have direct control over patrolling decisions, the same cannot be said for the ethnic composition of suspects they encounter. Thus, analysing patrolling decisions allows us to decompose over-searching into officer- and force-level decision making.

Our analysis so far treats the officer patrolling areas as given. However, patrolling areas are not allocated at random. Rather, police departments' deployment decisions are the consequence of prioritising certain areas. Similarly to how we constructed an officer-specific measure of over-searching relative to patrol, we can construct a force-level measure of over-searching relative to population share. This allows us to multiplicatively decompose force-level over-searching into three factors: officer over-searching, over-patrolling and the aggregation discrepancy between officer and force level.

$$\begin{aligned} \text{Force over-searching} &= \text{Officer over-searching} \times \text{Over-patrolling} \times \text{Aggregation discrepancy} \\ \frac{\text{Force search share}}{\text{Population share}} &= \frac{\text{Officer search share}}{\text{Officer patrol share}} \times \frac{\text{Officer patrol share}}{\text{Population share}} \times \frac{\text{Force search share}}{\text{Officer search share}} \end{aligned}$$

Officer over-searching is just O^P —our measure of officer over-searching by an officer relative to patrol. Over-patrolling is the disparity between the individual officer's patrol share and the population share in the police force area. Last, the aggregation discrepancy is the disparity between the force-level search share and the officer's individual search share. In some sense, the aggregation discrepancy is simply a mathematical artefact to allow for the decomposition. It expresses how different this officer's search share is from the overall force-level search share. As we will see below its distribution represents the variation of officer search shares in relation to the force-level aggregated search share.

For example, we can decompose the over-searching of Asian people based on the medians of these three terms. Relative to population, Asian people are over-searched at the force level by a factor of $0.2506/0.1982 \approx 1.26$, which is their share in all searches by the police force divided by their population share. Median officer over-searching is $0.2335/0.2304 \approx 0.99$ which means that the median police officer does not over-search Asian people relative to patrol. Median over-patrolling is $0.2304/0.1982 \approx 1.16$ meaning that the median officer over-patrols Asian communities by a factor of 1.16. The aggregation discrepancy is

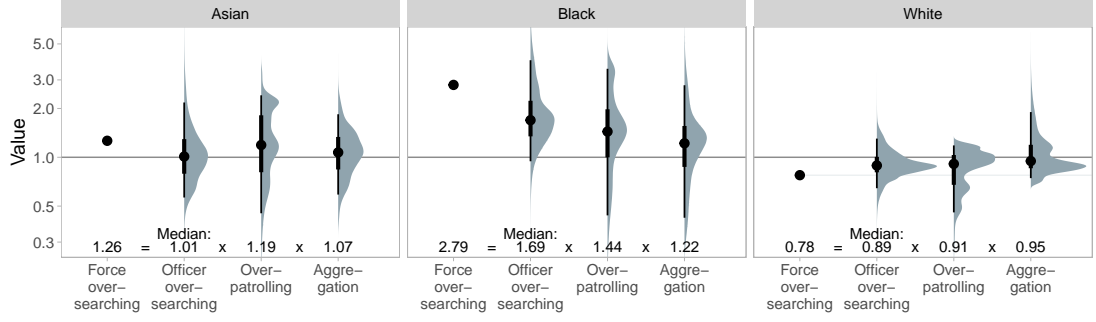


Figure 1.5: Decomposition of over-searching into officer over-searching, over-patrolling and aggregation discrepancy. Grey areas show posterior densities of terms calculated based on officers’ entire posterior distributions. Black dots represent the median of these densities and black lines represent 50% and 90% uncertainty intervals. The figure is based on 10,103 observations.

$0.2506/0.2335 \approx 1.07$ meaning that the median officer search share is slightly higher than the force-level average Asian search share.

Of course, any summary based on medians alone would be unsatisfactory. We therefore study the distributions over these three terms as induced by the officer-specific posteriors. On a practical level, this entails calculating them for every draw from each officer-specific posterior, the result of which is shown in Figure 1.5. Note that the distributions of officer over-searching shown in Figure 1.5 are the same as in Figure 1.3. At this point, it is important to recall that aggregated officer over-searching of Asian people obscures that some officers over- and some officers under-search Asian people relative to patrol which “cancels out” on the aggregate, resulting in a median of 1.01. This is only a concern for Asian over-searching since only there did the officer-level patterns differ from the aggregate. Taken together, the over-representation of Asian people in stop and search is accounted for by a combination of over-patrolling and some officer over-searching, though on aggregate officers do not over-search Asian people.

Black over-searching decomposes differently: Relative to the population Black people are over-searched at the force level by a factor of 2.79 which is primarily accounted for by officer over-searching. Still, over-patrolling also contributes to the overall over-searching of Black people. Last, we find that White people are under-searched at the force level. This is primarily accounted for by officers under-searching White people but also by under-patrolling of White areas. As we already saw in Figure 1.1, there is some variation between officers in their shares of White searches which is reflected in the aggregation discrepancy as officers’ search shares vary relative to the force-level average.

1.6 Discussion

Ethnic minorities are over-represented in stop and search compared to both their representation in the population and in crime. Our analysis exploits a panel of officers' searches from a major police force in England. We investigate the role of individual officers and police structures in the over-searching of ethnic minorities in stop and search.

For each officer, we first infer officer-specific search shares—the share of an ethnic group in an officer's searches. The ethnic composition of officers' searches is not meaningfully explained by officer characteristics. For example, an officer's ethnicity is not associated with the officer's searches, which matches some the mixed literature on the effect of officer ethnicity on policing outcomes (Mastrofski, Parks, and Worden, 1998; Engel, Sobol, and Worden, 2000; Terrill, 2001; Antonovics and Knight, 2009) and differs from some of it (McCrary, 2007; Legewie and Fagan, 2016). Instead, the ethnic compositions of officers' crime suspect encounters (suspect share) and of the officers' patrolling areas (patrol share) are associated with the ethnic composition of searches.

In exploring team compositions, we uncover a nuanced role of officer ethnicity. We find that teams' ethnic compositions are associated with officers' search behaviour: Teams that are more homogeneously White have lower minority search shares. Officers preferring to interact (or being tasked with interacting) with members of their own ethnicity alone cannot explain this association because more diverse teams search more Black people, yet most of this diversity is due to Asian officers and not Black officers, of whom there are very few. Instead, stereotype threat could explain why homogeneously White teams are associated with fewer searches of ethnic minorities as predominantly White teams might feel stereotyped as racist and avoid searches of minorities (Goff, Steele, and Davies, 2008).

In a second step, we infer an officer's bias of over-searching an ethnic group relative to crime suspects or to patrol. Almost all officers over-search Black people both relative to how often they encounter them as crime suspects and relative to the areas officers patrol. Similarly, almost all officers under-search White people relative to crime suspects or to patrol. For Asian people, we find that almost all officers search Asian people more than they encounter them as crime suspects. Relative to their patrol areas however, the picture shifts and officers are split into two groups, one that over-searches and another that under-searches Asian people which cancels out on the aggregate. One possible explanation for the split might

be the pooling of diverse ethnic identities of people with Indian, Bangladeshi or Pakistani backgrounds into a singular ‘Asian’ group. Search rates are not equal for these different groups and individuals with Indian backgrounds are searched at lower rates (Home Office, 2018a). The split of officer over-searching relative to patrol might then be an artefact of this pooling of ethnic identities.

Such disproportionate contact with police relates back to use of force. Ross, Winterhalder, and McElreath (2018) demonstrate that pooled analyses of use of force conditioned on the rates with which police encounter civilians can be confounded if officers differ in how often they encounter minorities. We find that officers indeed differ in how often they come into contact with ethnic minorities (for example, by searching them) and this cannot be explained by differential crime rates. Furthermore, even if officers were to use force on ethnic groups equally conditional on coming into contact with them, the fact that they have more contact with ethnic minorities means that these groups are subjected to higher levels of police use of force (Eckhouse, 2018). Of course, this not only applies to use of force but also other policing activities such as misdemeanour enforcement or arrests and emphasises the importance of documenting these disparities.

Regarding our findings of over-searching minorities relative to patrol, it is important to note that the patrol share is based on residential data from the 2011 ONS Census. The population available on the street, the ‘available population’, can be markedly different from the residential population (Malleon and Andresen, 2015). In particular, the ethnic make up of the available population can be different from the residential population and potentially account for the bias against ethnic minorities (Miller, Le Masurier, and Wicks, 2000; Waddington, Stenson, and Don, 2004). On the other hand, the available population explanation can be another self-fulfilling prophecy similar to the crime explanation (Equality and Human Rights Commission, 2010; Shiner and Delsol, 2015): If officers are deployed to areas with ethnically diverse available populations then the available population will predictably ‘explain away’ the bias compared to the residential population. That does not make the deployment decision bias-free. Other studies suggest that other area features such as its affluence also influence officers’ readiness to initiate searches (Shiner, Carre, et al., 2018). Search decisions have to be based on sufficient groups that a specific person is suspicious, not general availability of an ethnic group or general features of the area (Delsol and Shiner, 2006; Bowling and Phillips, 2007).

Deployment decisions are relevant to our analysis. Minority communities

are over-patrolled: The median officer patrols an area which is 1.16 times more Asian and 1.37 times more Black than all of West Midlands. The overall over-representation of ethnic minorities in stop and search decomposes into officer bias and over-patrolling. With officers over-searching minorities and command deploying officers to more diverse areas, the effects of officer biases are exacerbated by these deployment decisions. This results in more over-searching of minorities than can be attributed to officer biases alone.

The over-policing of minority communities documented in our study is supported by a wide range of other studies finding the same phenomenon (Sampson and Lauritsen, 1997; Fagan and Davies, 2000; Kane, 2003; Whitfield, 2004; Haining and Law, 2007; Williams, 2018; Otoyoy, 2018; Fatsis, 2019). Addressing the common question if these deployment biases can be explained by crime patterns is difficult. By their presence in an area, police are more likely to observe and record crime there. The observation of crime then is not independent from patrolling and searching patterns (and the ethnic biases therein). With the data available to us we cannot make any statement as to the mechanism that causes minority areas to be over-patrolled or the role of crime in that. Here, we only note that over-patrolling accounts for a considerable part of the overall over-searching of ethnic minorities.

There are clear limitations to our analysis, especially related to the generality of our findings. The policing context in the United Kingdom is particular, due to public and political scrutiny of police forces and the specific nature of the relationship between minority communities and the police. More officer-level analyses are needed and we hope that more police forces make officer-level data available to researchers. Furthermore, we hope that future work can clarify the process of deployment decisions.

For policy-makers, police forces and advocates looking to address the over-representation of ethnic minorities in stop and search, our results are both concerning and promising. Concerning, because our results show that 1. officer bias is a key factor in the over-representation of ethnic minorities in stop and search and 2. this officer bias is exacerbated by where police officers are deployed to. Promising, because our results could indicate a multiplier effect of institutional change where a reduction in anti-Asian and anti-Black bias in the police force applies both the searching officers on the street and to the officers making deployment decisions. Clearly, police forces should carefully examine their deployment policies as an amplifier in the over-representation of ethnic minorities in stop and search. Additionally though, we find that teams' ethnic compositions impact the

composition of officers' searches. Addressing the norms and environment of officer teams could then change officers' behaviour rather than just reduce its effect (Shiner and Thornbury, 2019). Our work shows that police forces need to reconcile the joint role of officer behaviour and department-level decisions in ethnic disparities in stop and search.

Appendix 1.A Sample selection

The final data set is compiled from four data files provided by West Midlands Police: `crimes`, `incidents`, `searches` and `officers`. The `searches` data are based on search forms which each officer has to fill out at the time of search. The searched person is then provided with a receipt and reference number of this record. We assign search decisions to all officers who jointly made the decision on patrol together, independently of who logged the search.

We link officers to crime suspects using the `incidents` and `crimes` data. Officers attend incidents throughout their work day. Some of these incidents will be logged as crimes and the `crimes` data holds information on the person suspected of having committed the crime. If an incident with officers A and B present is logged as a crime with suspect C present, we say that both officers A and B interacted with suspect C. We cannot ascertain whether suspect C was identified at the time of the crime incident or later on following an investigation. We exclude all crimes with more than five years between the crime incident and the crime report since it is unlikely that the officers encountered C as part of their investigation. All exclusions and matches between the data files are reported in Table 1.2. Our reliance on the `incidents` data to match officers to crime cases means that our final data does not contain any crimes which were reported at police stations. Our analysis also excludes all crimes which were recorded as a consequence of a stop and search. This means that our measure of the criminal population is not confounded by the process of stop and search.

In our analysis, we compare stop and searches between groups who self-identify as Asian, Black or White. During any interaction with police, individuals are asked to define their ethnicity into five broad categories: White, Mixed, Asian/Asian British, Black/Black British and Other. The White category encompasses encompasses British White, Irish and any other White background; the Asian/Asian British category encompasses Indian, Pakistani, Bangladeshi and any other Asian background and the Black/Black British category encompasses

Data set	Number of cases
incidents and crimes	
total incidents between 01/04/2014–30/09/2018	2,315,348
resulting in crime report	598,837
with any crime suspect information	341,297
with Asian, Black or White suspect	313,365
excluding old cases	312,651
by qualifying officers	
searches	
total stops and searches in study period	62,804
with stopped person’s ethnicity	59,739
with Asian, Black or White stopped person	56,021
requiring reasonable grounds of suspicion	55,740
by qualifying officers	36,028
officers	
total active police officers in West Midlands	5,081
were active in the police force in at least 5 out of 9 half-years	3,916
performed at least one search	1,194

Table 1.2: Details of matching and exclusion criteria applied to **incidents**, **crimes**, **stops** and **officers**. Indented conditions are chained: the last row of this table are all officers who were active in at least 5 out of 9 half-years AND performed at least one search in that time.

Caribbean, African and any other Black background. This classification system used by the police is based on the ONS 2001 Census (Office for National Statistics, 2003; Bowsher, 2007).

In the ONS 2011 Census, the Office for National Statistics changed the classification system to include Chinese people in the Asian/Asian British category rather than in the Other code as they did in 2001 (Office for National Statistics, 2009). To harmonise the ONS and the police’s classification system, we follow the police’s classification and exclude Chinese people from the census counts of Asian people.

Appendix 1.B Additional model results

In this section, we provide some additional results which do not currently have a place in the main text but may be of interest to the reader. We include the distributions presented in Figures 1.1 and 1.3 broken down by time interval in Figures 1.6 and 1.7. We also demonstrate the model fit in Figure 1.8. Table 1.3

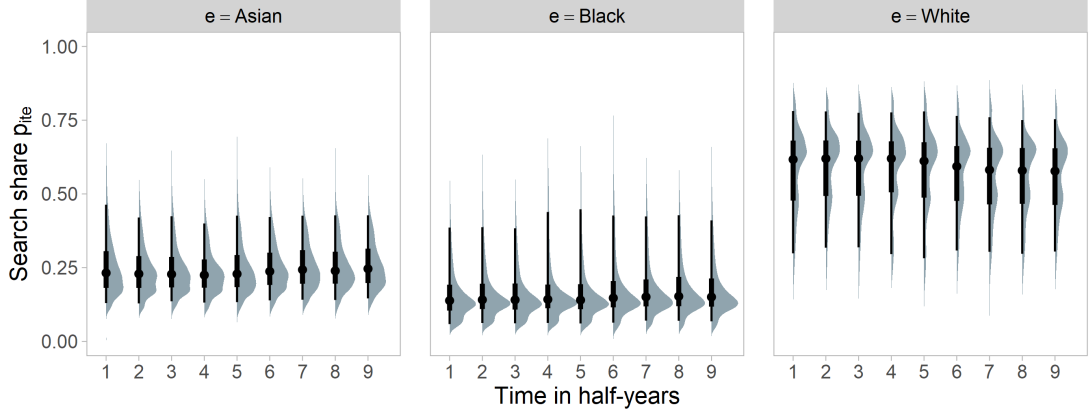


Figure 1.6: Disaggregated densities of posterior distributions of search shares over all officers for each 6-month time period based on 10,103 observations of officer search counts. The black dot represents the median of the distributions aggregated over officers and the black lines show 50% and 90% uncertainty intervals.

gives medians and 90% UI of the posterior distributions for the coefficients shown in Figure 1.2 in tabular format.

Parameter	Asian		Black	
	Median	90% UI	Median	90% UI
Global intercept	-0.09	[-1.25, 1.10]	-0.10	[-1.31, 1.31]
Female officer	-0.16	[-0.21, -0.10]	-0.09	[-0.15, -0.03]
Officer age	0.02	[-0.00, 0.04]	0.01	[-0.01, 0.04]
Officer experience	-0.03	[-0.05, -0.01]	-0.06	[-0.09, -0.04]
Officer of same ethnicity	-0.03	[-0.08, 0.03]	0.07	[-0.17, 0.29]
White share in team	-1.77	[-3.08, -0.50]	-2.18	[-3.71, -0.85]
Officer-level suspect share	1.95	[1.48, 2.35]	-2.29	[-2.67, -1.86]
Officer-level patrol share	1.86	[1.56, 2.20]	10.11	[9.50, 10.71]
SD of team-specific intercept (σ_α)	0.35	[0.27, 0.46]	0.43	[0.34, 0.57]

Table 1.3: Estimates and 90% uncertainty intervals (UI) for model parameters in Equation (1.2). The estimates are also displayed graphically in Figure 1.2. Officer age and experience are standardised.

Appendix 1.C AR(1) model

Lastly, we comment on the autocorrelation or serial correlation of O^S and O^P over time. If an officer exhibited a similar degree of bias against an ethnic group at all

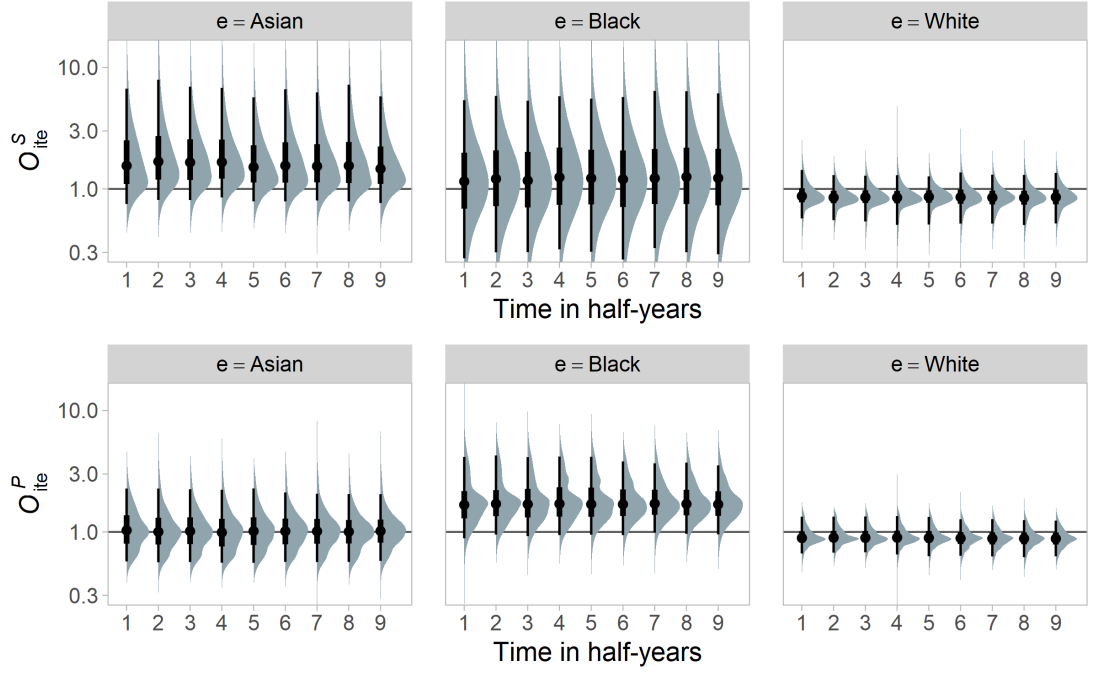


Figure 1.7: Posterior densities of O^S and O^P disaggregated by time period based on 10,103 observations of officer search counts. For visual clarity, we only show values between $[0.3, 10.0]$. Note that the y-axis is on the log scale. The black dots represent the medians; the black lines represent 50% and 90% uncertainty intervals.

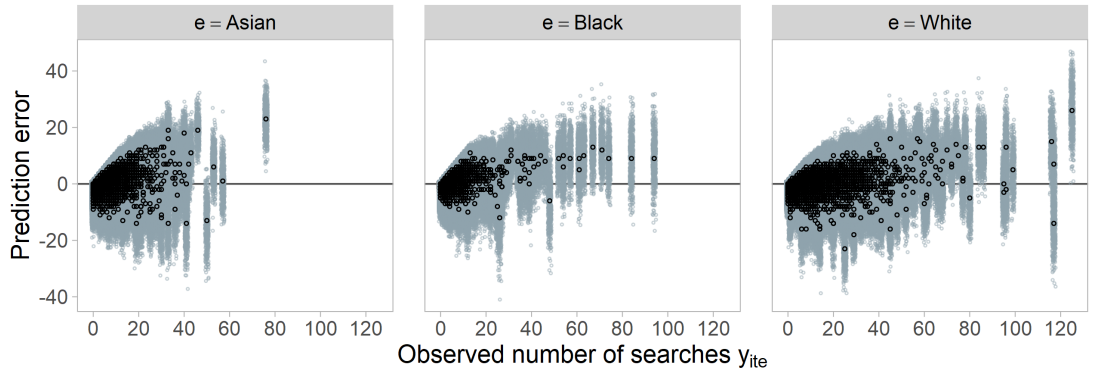


Figure 1.8: Comparison of observed search counts to predicted search counts based on inferred search shares p . Each grey dot is the observed search count by an officer in time period t and ethnic group e against the prediction error (observed – predicted). The black dots show the observed data against the error from the median prediction for that observation. The plot shows that key features of the data are captured in the model. The figure is based on 10,103 observations of officer search counts.

times, then we would observe a high degree of autocorrelation. Equally, if an officer's bias at a previous period does not give us any information about the officer's bias now then the bias is not stable and we would observe no autocorrelation.

We infer the officers' degree of autocorrelation using a autoregressive time series model of order 1, an AR(1) model. Briefly, an AR(1) model is a linear model that predicts the value of the time series at time t using the previous value of the series at $t-1$. The coefficient b_{ie} on the previous value gives us the degree of autocorrelation. For each officer and ethnicity we infer a separate coefficient such that we account for different degrees of autocorrelation between ethnic groups within the same officer.

In a second step, we model the stability of our over-searching measures O because it allows us to draw conclusions about the stability of officer bias. For each officer we obtain a posterior distribution over O_{ite} which is officer i 's log disparity of searching ethnicity e in time period t relative to ethnicity e 's prevalence in the officer's baseline. We then fit an autoregressive time series process of order 1, an AR(1) process to the time series of summarised O_{ite} over t . Since we cannot fit a time series to every single posterior draw we instead fit the time series at three summary points of the posterior distributions of O_{ite}^S and O_{ite}^P : The median and the lower and upper 90% uncertainty intervals. We consider two measures, O^S and O^P and we fit AR(1) models to both time series at three summary points. This results in $1,194 \text{ officers} \times 3 \text{ ethnic groups} \times 2 \text{ disparity measures} \times 3 \text{ summary points} = 14,316$ AR(1) coefficients. For ease of notation, we describe our model with respect to a generic over-searching measure O :

$$\begin{aligned} O_{ite} &= a_{ie} + b_{ie}O_{i(t-1)e} + \varepsilon_{ite}, & t > 1 \\ \varepsilon_{ite} &\sim N(0, \sigma_{ie}) \end{aligned}$$

where b_{ie} gives the degree of autocorrelation. If $b_{ie} > 0$, i.e., the autocorrelation is positive, then the O_{ite} move in the same direction over time. Negative autocorrelation indicates that the terms move in opposite directions over time. If b_{ie} is zero then the process is driven entirely by a_{ie} and the error term.

Our time series is very short with only nine time periods. Additionally, some officers are not observed in the entire study period so we have even fewer observations for these officers. Altogether, the data sparsity makes the estimation of the officer-specific terms a_{ie} , b_{ie} and σ_{ie} challenging. We therefore introduce a hierarchical prior structure where all officer- and ethnicity-specific intercept and

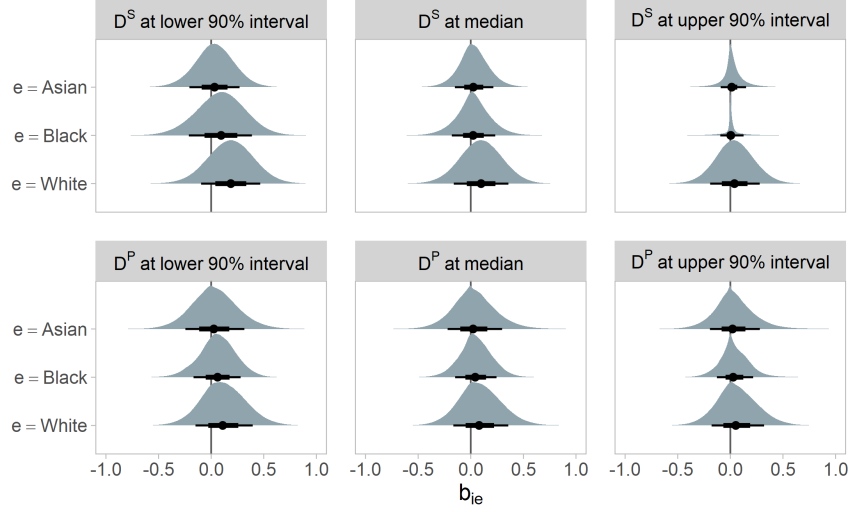


Figure 1.9: Densities of AR(1) coefficients for time series of O^S and O^P at different summary points of the posterior distributions of O_{ite}^S and O_{ite}^P . Each panel of this figure is based on 4,000 posterior draws each from the posterior distributions of $1,194 \times 3 = 3,582$ inferred AR(1) coefficients.

slope a and b have an ethnicity-specific hyper-prior A_e or B_e on their mean. Additionally, the standard deviation σ_{ie} also has an ethnicity-specific hyper-prior σ_e . More formally:

$$\begin{aligned} a_{ie} &\sim N(A_e, 0.25) & A_e &\sim N(0, 1) \\ b_{ie} &\sim N(B_e, 0.25) & B_e &\sim N(0, 0.25) \\ \sigma_{ie} &\sim N^+(0, \sigma_e) & \sigma_e &\sim N^+(0, 1), \end{aligned}$$

where N^+ denotes a half-normal distribution.

Unfortunately, our time series of only nine half-years is too short to allow strong conclusions about the stability of bias. Figure 1.9 shows the distributions of AR(1) coefficients aggregated by officers at the different summary points. The 90% uncertainty intervals around the AR(1) coefficients are simply too wide: the average range between the lower end of the 90% UI and the upper end of the 90% is 0.66 which is considerable given that the coefficient lies between -1 and 1.

Evidence of autocorrelation is particularly weak for O^S , where the 90% UI for only 56 of the 3579 coefficients excluded zero. In contrast, 1,355 out of 3579 90% UI for the coefficients estimating the stability of O^P excluded zero. All coefficients are positive. In other words, we have weak evidence that approximately one third of officers are consistent in their search bias against ethnic groups relative to their patrol baseline. We cannot comment on the strength of this consistency because

of the aforementioned poor estimation of the coefficients. We simply do not have enough data to make conclusive statements about the stability of bias for the majority of officers.

Chapter 2

(No) Spillovers in reporting domestic abuse to police

2.1 Introduction

Immediately after a crime, the risk of another crime in the same neighbourhood is increased (Johnson and Bowers, 2004; Short et al., 2009; Mohler et al., 2011). This triggering behaviour has been documented for a diverse range of crimes such as burglaries, homicides and robberies (Mohler, 2014; Reinhart and Greenhouse, 2018; Flaxman et al., 2019). Typically, the triggering is due to the same offender repeating their crime or other offenders learning about the promising criminal opportunity (Bernasco, 2008). However, this focus on offender behaviour as a source of triggering potentially overlooks victim behaviour as a source. Crime victims share their experiences with crime and with reporting to the police. This could influence other crime victims in their reporting decisions and imply triggering of crime reporting.

The reporting of crime depends on many factors which not only have to do with the characteristics of the crime, but with the geographic and social context of the crime victim (Bachman and Coker, 1995; Goudriaan, Wittebrood, and Nieuwbeerta, 2006). Studies have shown that some behaviours exhibit triggering behaviour: The behaviour of one person can spillover into their immediate social and spatial environment (e.g., Bayer, Hjalmarsson, and Pozen, 2009; Fadlon and Nielsen, 2019). Combined with the well-documented triggering behaviour of crime on the offender side, we hypothesize that this property extends to the *reporting* of crime to police by victims. We explore this question in the context of domestic abuse.

Domestic abuse is a particularly interesting context to study spillovers effects due to four facts: 1. It is highly prevalent, 2. significantly under-reported, 3. has a high degree of social disclosure, that is many victims/survivors¹ tell a friend, neighbour or family member about the abuse and 4. typically, perpetrator and victim(s) constellations remain stable over time such that one perpetrator acts within only one household. That means that if we were to document any spillovers in reports of domestic abuse, they must be due to spillovers in victim behaviour rather than in offender behaviour.

The prevalence of domestic abuse is disconcerting: In the United Kingdom, approximately one in four adults experience domestic abuse in their adult life and an estimated 2.4 million adults have experienced some form of it in the last year (Office for National Statistics, 2019a). Domestic abuse is a complex phenomenon consisting of abusive events as well as patterns of abusive behaviours at the hands of intimate partners or family members (Kelly and Westmorland, 2016). At the same time, domestic abuse is largely hidden from public view. It is one of the most under-reported serious crimes and victims/survivors often endure abuse for many years before seeking outside institutional support with police or a support provider (SafeLives, 2015). Reporting a crime to police is a complex choice, especially in the context of domestic abuse. At the same time, the majority of survivors disclose the abuse to a friend, neighbour or family member, at a much higher rate than reporting to police (Osborne, Lau, and Britton, 2012).

High rates of social disclosure paired with the high incidence of domestic abuse provide a context in which the disclosure of domestic abuse to police by *others* could affect an individual's decision to report. However, identifying spillovers in domestic abuse reporting is challenging because the number of calls to police about domestic abuse varies, both due to the variation in the underlying abuse and due to variation in reporting behaviour (Cohn, 1993). Furthermore, since neighbourhood characteristics have been shown to influence both the incidence of domestic abuse and the reporting of crime, we would expect both to cluster in neighbourhoods (Goudriaan, Wittebrood, and Nieuwbeerta, 2006; Beyer, Wallis, and Hamberger, 2015).

¹Often the term victim is used to describe someone who experienced domestic abuse recently, or in the criminal justice context. Some have criticized the term for assigning the person experiencing domestic abuse a passive role and prefer the term survivor. It emphasizes individuals' agency in processing and recovering from the abuse. Some people who experienced domestic abuse prefer to identify as victim, while others prefer survivor. In this chapter we use both terms interchangeably to reflect the tension between the victimisation that people experiencing abuse go through and the agency they possess to define, interpret and deal with the experience on their own terms.

To separate spillover effects from spatio-temporal clustering, we model a data set of domestic abuse reports to police using a Hawkes process (Hawkes, 1971; Ogata, 1988; Reinhart, 2018). Hawkes-type triggering spatio-temporal point processes have been an invaluable tool in identifying spillovers in criminal behaviour and disentangling them from clusters in space and time (Mohler, 2014; Reinhart and Greenhouse, 2018; Flaxman et al., 2019). The Hawkes process predicts a series of discrete events (reports) that arise from one of two intensity components: the background component which corresponds to the “typical” incidence of domestic abuse reports or the triggering component where each past event can trigger the occurrence of future events. The Hawkes process does not identify a causal link between events, as in “event i caused event j ”. Instead, past events increase the likelihood of future events in their spatio-temporal vicinity. Triggering is therefore the statistical quantification of said increase or spillover, not the identification of a causal relationship.

Specifying the forms of the background and triggering components is a crucial modelling choice since their shapes have important implications for inference and detection of triggering behaviour (Reinhart and Greenhouse, 2018). We employ the specification of the Hawkes process proposed by Zhuang and Mateu (2019) because it includes periodic components that account for the daily and weekly periodicity of reporting behaviour, for example if many domestic abuse reports are made Friday nights. Periodicity is predicted by criminological theories such as routine activity as well as empirical data (Cohen and Felson, 1979; Rotton and Cohn, 2001; Johnson and Bowers, 2004).

We also develop an extension of Zhuang and Mateu (2019)’s model such that reports of domestic abuse can be triggered by past reports, but also by the police response to past events. A meta-study by Davis, Weisburd, and Taylor (2008) analysed the effect of police returning to the incident household to provide support to the initial responding officers or to follow-up with the victim(s). They find that such follow-up visits do not reduce future violence occurring but increase victim reporting of the violence. It stands to reason that if follow-up visits by police amplify the effects of a report to police, this effect could potentially extend beyond the reporting household. Our model extension adding an additional spillover channel tests this explicitly. With this extension that allows two types of events to trigger reports of domestic abuse (reports and follow-ups), we can statistically distinguish two channels of spillover effects: 1. report-to-report and 2. follow-up-to-report spillovers.

We test the presence of spillover effects in domestic abuse reporting on a

data set of 6,084 calls for service to police. The data set covers all calls flagged as concerning domestic abuse in a major English city between January 2018 to December 2018. We find extremely limited evidence of spillover effects of domestic abuse reporting. Any such effects are limited to the first 6 days after an initial report and an area of 400m around the event. This is true for both types of spillovers tested in this study, report-to-report and follow-up-to-report. We find weak evidence that events taking place in neighbourhoods where people live closer together increase the likelihood of future events more than events in less dense areas. Taken together however, triggering in domestic abuse reporting is negligible.

Instead we document highly periodic reporting behaviour. In line with other studies, reports of domestic abuse are highest on weekends and in the evening. Our study highlights that when modelling crime contagion, it is important to carefully think about the dynamics of the reporting behaviour underlying the data.

The remainder of the chapter is structured as follows: Section 2.2 discusses the context of spillovers in reporting. Sections 2.3 and 2.4 introduce the data and methodology used in this study. Section 2.5 presents our results and we conclude by discussing the implications in Section 2.6.

2.2 Motivation

The focus of our study is on potential spillover effects of domestic abuse reporting. Section 2.2.1 discusses spillover effects in criminal behaviour due to offender behaviour identified using Hawkes processes. Section 2.2.2 examines the victim's decision to report domestic abuse to police. Section 2.2.3 concludes by discussing why spillover effects might exist in domestic abuse reporting and potential mechanisms. A full discussion of why we did not find any empirical evidence supporting this is postponed until Section 2.6.

2.2.1 Spillovers in crime

Spatio-temporal point processes are stochastic processes which model the occurrence of discrete events in space and time. In this study, we employ a spatio-temporal Hawkes-type point process with a triggering (or self-exciting) component (Hawkes, 1971; Ogata, 1988). Originally popular to model aftershocks of earthquakes, Hawkes processes have also been used to model crime events such

as burglaries, shootings and calls for service more generally (Mohler et al., 2011; Reinhart and Greenhouse, 2018; Loeffler and Flaxman, 2018; Flaxman et al., 2019; Zhuang and Mateu, 2019). Their use is motivated by the observation that one crime tends to trigger further crimes in the same area immediately afterwards. For example, burglars often re-visit an area in the weeks after a successful score (Short et al., 2009; Bernasco, Johnson, and Ruiter, 2015). Such behaviour is very conveniently modelled by a Hawkes process because it can disentangle clusters from spillovers. Put differently, is the risk of burglary in an area high because few houses have alarm systems or because of a recent break-in? The Hawkes process can be understood as a model-based test of this distinction, in contrast with traditional statistics vulnerable to arbitrary thresholds (Meyer et al., 2016; Loeffler and Flaxman, 2018).

The triggering behaviour of crime has been described under different names: Near-repeat victimization describes the phenomenon that a small number of victims or victims with similar characteristics account for a large number of crime offences (Farrell and Pease, 1993). Weisburd (2015) argues for a ‘law of crime concentration’ which states that just a few street segments account for the vast majority of crimes in a city. Such observations have led to a rich literature of the identification of crime hotspots, both stable and emerging in time (Johnson and Bowers, 2004; Gorr and Lee, 2015).

Multiple reasons have been put forward as to why crime exhibits triggering behaviour. Criminological theories such as routine activity theory analyse crimes as the intersection of a suitable target or victim, a motivated offender and a lack of supervision (Cohen and Felson, 1979). Similarly, the application of an economic or rational choice framework to crime predicts that a rational decision-maker will consider potential costs and payoffs of a criminal opportunity (Clarke and Cornish, 1985; Sanders, Kuhns, and Blevins, 2017). Under both theories, a successful offender will seek to repeat their success in addition to other offenders picking up similar cues (Bernasco, 2008). Lastly, some crimes will result in retaliatory action. For example, gang violence can induce retaliatory violence, as can shootings (Ratcliffe and Rengert, 2008; Brantingham et al., 2018). These behaviours have been successfully identified using Hawkes processes (Mohler et al., 2011; Reinhart and Greenhouse, 2018; Loeffler and Flaxman, 2018).

But this research focuses only on the offender side of crime. A phenomenon that has not yet been explored in depth is whether there are any spillovers in victim behaviour. The context in which victims of a crime make the decision to

report to police is explored in more detail in the next section. But a rich literature of behaviour in other contexts provide us with a reference frame on how crime victim behaviour might spill over: There is ample evidence that a change in a person’s or household’s behaviour can change the behaviour of surrounding people and households. For example, sending letters about TV licenses to households increases compliance in non-treated households in the neighbourhood (Rincke and Traxler, 2011; Drago, Mengel, and Traxler, 2020). Beyond license compliance, such effects have been documented in, e.g., voting, insurance, or school performance (Nickerson, 2008; Hong and Raudenbush, 2006; Sobel, 2006; Cai, De Janvry, and Sadoulet, 2015; Halloran and Hudgens, 2016). Identifying spillovers is relatively feasible when evaluating a specific, randomized intervention (Aronow et al., 2020).

In observational studies however, one encounters a familiar issue: Is the similarity in behaviour the result of spillover or simply a concentration of behaviour? For example, Bertrand, Luttmer, and Mullainathan (2000) demonstrate that women are more likely to use welfare when there is a local network of the same ethnic group (and the group also has a high level of welfare use). Aizer and Currie (2004) try to separate neighbourhood from network effects in the similar use of publicly funded prenatal care within ethnic groups. In contrast, they find no evidence for information sharing through networks and instead show that behaviour is highly similar in ethnic groups because of local hospital policies.

Without assumptions or explicit knowledge about the underlying structure of networks (e.g., Fadlon and Nielsen, 2019; Nicoletti, Salvanes, and Tominey, 2018) it is challenging to separate clusters and spillovers. It is in this precise context that the Hawkes process becomes a particularly valuable statistical model. Meyer et al. (2016) argue that Hawkes processes are a principled, model-based way of separating space-time clusters from the spread of a behaviour.

As we will argue in the next section, domestic abuse is an ideal context in which to explore the potential for reporting spillovers: It is under-reported but highly prevalent and exhibits a high degree of social disclosure while typically only having one offender/victim constellation per household. In our context, the use of a Hawkes process allows us to identify if a single report of domestic abuse to police increases the likelihood that a victim of domestic abuse in another household in the neighbourhood will also report without having to make any assumptions about the nature or structure of local networks.

2.2.2 Reporting domestic abuse

A persistent challenge to understanding and addressing domestic abuse is its hidden nature. It is hidden because the abuse often takes place away from the public eye in private homes, but also because domestic abuse is extremely under-reported. Many victims endure domestic abuse for a long period before disclosing their abuse to a formal institution, if at all (SafeLives, 2015). Because of this, available numbers are significant undercounts of the actual prevalence. Even large-scale surveys of the general population such as the Crime Survey for England and Wales are highly sensitive to methodological details (Ellsberg, Heise, et al., 2001; Walby and Allen, 2004; Emery, 2010; Agüero and Frisanco, 2021). Further, they are most likely underestimates of the true prevalence because they often exclude people outside of stable households such as unhoused individuals or those in temporary accommodation, hospitals and refuges (Office for National Statistics, 2017).

Given its high prevalence, it is crucial to consider why victims/survivors of domestic abuse do not report to police in higher numbers (Osborne, Lau, and Britton, 2012). In general, the decision to report a crime to the police is framed as a process of weighing potential positive outcomes to reporting against disincentives (Laub, 1981; Skogan, 1984; Gottfredson and Gottfredson, 1988). Both of these factors are particularly fraught within the context of domestic abuse since victims face significant barriers to reporting and judicial outcomes are usually poor.

Domestic abuse cases have high rates of attrition through the various stages of criminal justice system such that 96% of police-recorded cases do not result in a conviction (Hester, 2006; Her Majesty’s Inspectorate of Constabulary, 2014). Agents within the system often attribute this to victims withdrawing participation (Hester, 2006; Starmer, 2011). In contrast, other work has identified insufficient evidence collection and under-charging as a significant factor in attrition (Nelson, 2013; Her Majesty’s Inspectorate of Constabulary, 2014). Interviews of survivors find that participation in the criminal justice system process for them is highly dependent on their perception of the system’s ability to provide safety (Felson et al., 2002; Hester, 2006). These interviews emphasize that justice goals of victims/survivors can differ from those of the criminal justice system and provide an important explanation for under-reporting (Coy and Kelly, 2011; Westmarland and Kelly, 2013).

Studies examining the challenges to reporting domestic abuse have identified

a number of barriers. A crucial first step is recognizing the abuse as a serious offence worth reporting. Often, domestic abuse is considered a “private issue” which should not be publicized to the outside (Tjaden and Thoennes, 2000b; Rogers et al., 2016). The recognition of abuse as such is strongly mediated by the seriousness of the abuse, where victims with more serious injuries are more likely to call the police (Bachman and Coker, 1995). Victims of other types of crimes are less likely to call the police if the offender is not a stranger (Gartner and Macmillan, 1995). Since domestic abuse by definition implies an offender intimately familiar to the victim, victims have to make a difficult choice of “handing over” a person close to them to the police. Involving the police can set into motion a series of consequences which victims may not necessarily want. For example, victims do not always want to leave the relationship or family environment due to love or family bonds (Strube, 1988). More importantly, many victims are potentially isolated or without means to leave the abusive environment. Combined with underfunded domestic abuse services, the decision to disclose domestic abuse can be existential (Walby and Towers, 2012; Sanders-McDonagh, Neville, and Nolas, 2016). Indeed, 10–40% of unhoused individuals cite domestic abuse as contributing factor to their homelessness (Cramer and Carter, 2002; Office for National Statistics, 2019b). Furthermore, victims of domestic abuse often fear retribution by the perpetrator or people close to the perpetrator and fear for the safety of their children (Strube, 1988; Greenfeld et al., 1998; Coy and Kelly, 2011).

Another important consideration is how perceptions of what constitutes “legitimate” abuse mediate reporting decisions. In the context of sexual assault, existing work has uncovered that notions of what a “real assault” looks like (e.g., a violent assault by an armed stranger) significantly affect victims’ willingness to report to police (Myhill and Allen, 2002). Recent work finds that perceptions of what constitutes sexual harassment is mediated by how stereotypically feminine the female victim presents (Goh et al., 2021). While this is less well-explored in the context of domestic abuse, characteristics of the abuse and its survivor matter: For example, intoxicated victims are assigned more blame for the abuse than sober victims (Leonard, 2001).

Such perceptions of “real” or “legitimate” abuse affect police officers as well. Officers, too, operate with beliefs and stereotypes of “typical victims” of domestic abuse and sexual violence, which influence how they handle these cases (Trute, Adkins, and MacDonald, 1992; Robinson, Pinchevsky, and Guthrie, 2018; O’Neal, 2019). Demographic groups such as sex workers, transgender people and

people with disabilities are already considerably more likely to be subjected to violence due to their marginalization but often face further challenges being taken seriously by police (Nixon, 2009; Roch, Ritchie, and Morton, 2010; Lombard and Scott, 2013; Phipps, 2013; Rogers et al., 2016). Some studies find that reports of domestic abuse increase as the number of female police officers in the force increases which suggests that some of these factors may be less pronounced in female officers (Miller and Segal, 2019; Kavanaugh, Sviatschi, and Trako, 2019).

This arbitration of legitimacy can turn police officers—who are often the first point of contact to the criminal justice system—into gatekeepers of access to said systems (Taylor and Gassner, 2010). Surveys have shown that victims fear not being believed or being taken seriously by police (Tjaden and Thoennes, 2000b; Hawkins and Laxton, 2014; Her Majesty’s Inspectorate of Constabulary, 2014). Indeed, there have been investigations into police mishandling of domestic abuse cases and into domestic abuse by police officers (Independent Office for Police Conduct, 2018; Centre for Women’s Justice, 2020). Similarly, victims/survivors who decide to report to police affirm that their experience is not uniformly positive. In the Crime Survey for England and Wales, 72% of victims of domestic abuse stated that they found the police fairly or very helpful, while only 55% reported feeling safer after contacting police. Approximately 14% reported feeling less safe (Osborne, Lau, and Britton, 2012). The upshot of analysing the factors influencing the decision to report domestic abuse to police is that many victims/survivors view calling the police as a last resort (Fitzgerald, Swan, and Fischer, 1995; Women’s Aid, 2009).

This would leave us with a scant premise for investigating spillovers of police reports of domestic abuse. However, surveys of survivors reveal that they do disclose their abuse, even if not necessarily to police: In England, more than 73% of victims of domestic abuse told a friend or relative about the abuse, compared with only 23% reporting to police (Osborne, Lau, and Britton, 2012). While the reporting rates to police vary (from as low as 2% to almost 50%), other work similarly finds that more than half of victims disclose their abuse to someone close to them (Greenberg and Ruback, 1992; Fisher et al., 2003; Coy and Kelly, 2011; Stark et al., 2013). Indeed, often the response by the confidant is influential in the victim’s decision to end the relationship as well as report to police (Goodkind et al., 2003; Regan et al., 2007; Biaggio, Brownell, and Watts, 1991).

High rates of social disclosure paired with the high incidence of domestic abuse provide a context in which the disclosure of domestic abuse to police by others could affect the decision to report. This establishes the basis for our core

hypothesis: Someone close to a victim of domestic abuse reporting their own abuse to police might induce the victim to contact police themselves.

2.2.3 Spillover channels

In our model, we distinguish between two channels of spillovers: Report-to-report and follow-up-to-report.

The first channel, report-to-report, accounts for spillovers due to information passing through social peer networks. As established, victims of domestic abuse often fail to identify criminal abuse as such because of their close relationship with the perpetrator. Knowing that others in their vicinity reported abuse might affect how victims perceive and frame their abuse. While this has not yet been explored in the context of domestic abuse, some studies have examined the effects of information transmission on the reporting of sexual violence: Cheng and Hsiaw (2020) develop a formal model of reporting sexual misconduct in the workplace. The context of workplace harassment differs from domestic abuse: Mainly, their work centres on corroboration, that is multiple individuals need to report misconduct before action against an offender is taken. As a consequence, individuals subjected to sexual harassment face strategic uncertainty and a coordination problem around reporting: If they report misconduct and no one else has, they may face retaliatory penalties. If instead multiple individuals have come forward and a substantive record can be corroborated, then an outside party can sanction the harasser. Corroboration is difficult to translate to the context of domestic abuse (where a perpetrator typically only abuses within their immediate household). Still, Cheng and Hsiaw (2020)’s model illustrates that individuals face information frictions about how wide-spread a behaviour is, which affects their propensity to report.

Levy and Mattsson (2020) study the effect of the #MeToo movement on reporting behaviour and find that it resulted in a persistent increase of reports of sexual violence. They argue that their results are plausibly explained by a rapid change in social norms and information. Similarly, Iyer et al. (2012) and McDougal et al. (2018) find that visible social changes (the election of female politicians and a highly publicized case of sexual violence) lead to a large increase of reports of sexual violence.

The second channel of spillovers, follow-up-to-report, accounts for the effects of police intervention. Intervention by police in cases of domestic abuse is

not uncontroversial: Historically, police departments tended to follow an under-enforcement policy which meant they rarely intervened and even less frequently arrested the perpetrator. Only after public pressure by activists did the policy response change (Fagan, 1996; Erwin, 2006). In the United States, police departments' non-arrest policies were subjects of lawsuits which argued that the (lack of) intervention did not provide women victims with equal protection of the law. As a result, many police departments in the United States implemented mandatory arrest policies, meaning that one person, the perpetrator, is to be arrested at the scene of the domestic abuse incident (Fagan, 1996). Multiple studies have since demonstrated that such mandatory arrest policies do not create deterrent effects and may lead to increased arrest rates of victims of abuse (Hoppe et al., 2020). Mandatory arrest policies were never formally implemented in the United Kingdom, but actively encouraged. Today, policing is an integral part of the United Kingdom's policy response to domestic abuse (Walklate, 2008; Matczak, Hatzidimitriadou, and Lindsay, 2011).

Some studies have investigated the effects of police intervention on future violence within the same household, with mixed results (Hanmer, Griffiths, and Jerwood, 1999; Hoppe et al., 2020). A meta analysis by Davis, Weisburd, and Taylor (2008) of ten studies investigates a specific type of police intervention: follow-up visits which they call "second responder visits". These in-person visits are part of specific programmes which aim to intervene in the cycle of domestic abuse. When victims notify police about an incident, a lot of time may pass before they call police again. This may be due to a cessation of the abuse or reflective of victims' reluctance to report to police. Crisis theory predicts that there may be a "window of opportunity" immediately after an incident of domestic abuse during which the victim might be interested in leaving the environment and/or pursue legal options because the usual coping strategies are not working (Kelly, Bindel, et al., 1999; Mickish, 2002). Therefore, the programmes evaluated in Davis, Weisburd, and Taylor (2008) send a police officer and, depending on the programme, police officer together with a victim advocate to follow up on the initial report within a few days.

Each study in the meta analysis considered recidivism as a primary outcome, with most studies using both police reports and victim surveys to measure the incidence of repeat violence. Half the studies were a fully randomized experimental design, half were quasi-experimental. All studies took place in the United States. Results from the victim surveys indicate that the follow-up visits had no significant effect on repeat violence (standardized difference in group means:

-0.01, $p = 0.82$). Instead, there is a modest positive increase in reports to police (standardized difference in group means: 0.12, $p = 0.01$). The results suggest that while the follow-up visits do not reduce the likelihood of abuse, victims seem more confident about reporting the violence to police.

Our study tests an extension of this effect: Do police visits following an incident have an effect outside the directly affected household? The mechanism of this effect links back to the notion that victims of domestic abuse often do not recognize their abuse as such. Repeat visits by police to another call can then serve as validation and amplification: Police are taking the incident seriously and paying attention.

We would expect any spillover effects to be more prevalent in areas with strong social networks. Unfortunately, we do not have any measure of local social cohesion available (as in, e.g., Goudriaan, Wittebrood, and Nieuwbeerta, 2006). As a proxy, we use the share of households living in detached houses in the area since individuals living closer together are also closer to the goings-on of their neighbours (see also Ivandic, Kirchmaier, and Linton, 2020).

2.3 Data

Our data set covers all 6,084 calls to police about an incident of domestic abuse between 01/01/2018 to 31/12/2018 in a city with over 300,000 inhabitants in England.

When someone calls the police for service, the call will be picked up by a call handler in the police contact centre. The handler will ask a series of questions to evaluate the situation and decide on an appropriate police response. If the call handler at this point assesses the situation to take place in the context of domestic abuse, he or she raises a flag in the system which notifies the responding officer of that context.

In the United Kingdom, there is no statutory crime of domestic abuse. But many forms of domestic violence constitute criminal offences such as assault, sexual offences, stalking or criminal damage. Police forces in the UK classify such incidents as domestic abuse if they meet the cross-government definition: “Any incident or pattern of incidents of controlling, coercive or threatening behaviour, violence or abuse between those aged 16 or over who are or have been intimate partners or family members regardless of gender or sexuality. This can encompass

but is not limited to the following types of abuse: psychological, physical, sexual, financial, emotional.” (Home Office, 2013).

In the past, English police forces have been criticized for not supplying responding officers with sufficient information. For example, officers may often not have any information on the perpetrator or know that the victim/survivor may be a repeat victim (Her Majesty’s Inspectorate of Constabulary, 2014). Similarly, the initial call handler may not identify a situation as domestic abuse but the police responders at the scene may do so.

A call enters our data set when either the handler or the responders classify the call as domestic abuse. Research conducted in the UK finds evidence that there is variation between call handlers and officers in their handling of domestic abuse: Female call handlers result in faster police response and cases handled by response teams with more female officers have lower legal attrition (Hawkins and Laxton, 2014).

For each call in our data set, we know the time when the call was placed and the location of the incident. A key feature of our analysis is the question of police follow-ups where officers return to places of domestic abuse. A key challenge here, however, is that we cannot consistently check *why* police officers return: Is it for a scheduled routine visit or is it because the domestic situation escalated and requires intervention? This reason for this inconsistency is inconsistent police record keeping. Some officers who return to the scene will link their new visit to the old case identifier which means that we can follow this link. However, some officers will also create a new case identifier unconnected to the original case. Is this because there is a new incident of domestic abuse at the house, therefore necessitating the creation of a new case file? Or is this simply an oversight on the officer’s part?

While more than two thirds of return visits happen between business hours, suggestive of scheduled visits, our approach to this issue is conservative: Every return visit by police officers to the same address within two weeks of the initial call is classified as a follow-up visit, without distinction as to why the officers might have returned. If officers return after more than 15 days of the initial call we consider it a new incident of domestic abuse due to escalated violence. This concerns only 29 calls in our data set. Choosing 15 days as the cutoff is motivated by two factors: 1. Police aim to respond to non-urgent incidents within 5 days (and within 15-60min to urgent incidents, depending on the urgency) which means that there is a reasonable range of days after an initial incident during which officers might follow-up and 2. our model uses an explicit cutoff at 30 days after

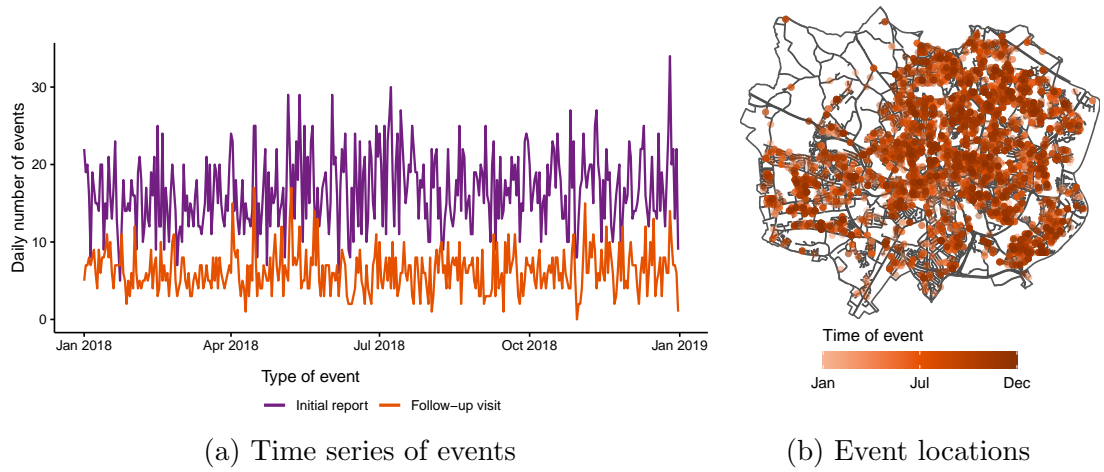


Figure 2.1: Time and location occurrence of events

which we no longer expect triggering of domestic abuse reports and classifying into this scheme implicitly prevents reports of domestic abuse in a household from apparently “triggering” further reports in the same household which are in reality the fallout from the initial report.

Taken together, this results in 6,084 initial calls for service and 2,286 follow-up visits by police. Figure 2.1a and Figure 2.1b show the temporal and spatial dimension of the raw data.

2.4 Method

We formally introduce our specification of the Hawkes process in Section 2.4.1 and our inference procedure in Section 2.4.2.

2.4.1 Model

We use a self-exciting point process model to describe reports of domestic abuse to police. Consider a set of observed realizations from a Hawkes process (Hawkes, 1971; Ogata, 1988), where $\{t_1, t_2, \dots, t_n\}$ denotes the time-ordered sequence of event times and $\{s_1, s_2, \dots, s_n\}$ denotes the time-ordered sequence of event locations. The conditional intensity of this spatio-temporal point process defined

on time $t \in [0, T)$ and location $s \in X \subseteq \mathbb{R}^d$ is then given by

$$\begin{aligned}\lambda(t, s | \mathcal{H}_t) &= \mu(t, s) + \int_0^t \int_X f(t - u, s - v) dN(u \times v) \\ &= \mu(t, s) + \sum_{j: t_j < t} f(t - t_j, s - s_j),\end{aligned}\tag{2.1}$$

where $N(\cdot)$ counts the number of events in an interval and $\mu(t, s)$ describes the background rate. Because of the self-exciting component f , the intensity is conditioned on the history of the process up to and including t , \mathcal{H}_t . The shape of f determines if, by how much and for how long past events can trigger events in addition to the base background rate. Because we need $\lambda(t, s | \mathcal{H}_t) \geq 0$, we set $\mu(t, s) \geq 0$ and $f(t, s) \geq 0$ for all t, s . For ease of notation, we will omit the explicit conditioning on \mathcal{H}_t from now on, but the reader should keep in mind that $f(t, s)$ depends on all past events \mathcal{H}_t , for all spatial locations s .

The specification of the background and triggering component depend on the application context (Reinhart, 2018). Our specification of the background component follows Zhuang and Mateu (2019), who consider a periodic decomposition of the background as follows:

$$\mu(t, s) = m_0 \mu_{\text{trend}}(t) \mu_{\text{weekly}}(t) \mu_{\text{daily}}(t) \mu_{\text{area}}(s),\tag{2.2}$$

where $\mu_{\text{trend}}(t)$, $\mu_{\text{weekly}}(t)$ and $\mu_{\text{daily}}(t)$ represent the trend term over the whole study window, the weekly and daily periodicity in the time dimension of the background rate, respectively. $\mu_{\text{area}}(s)$ is an estimate of the background spatial intensity in the study area. These terms are normalized to have mean 1. As a consequence, m_0 , which is a non-negative weighting term attains the role of weighting the entire background component (Loeffler and Flaxman, 2018; Zhuang and Mateu, 2019).

We take the triggering component f to be separable in time and space such that $f(t, s) = \theta g(t) h(s)$. Again we normalize g and h to integrate to 1 such that θ gives the average number of events coming from the trigger component. Furthermore, we extend the approach of Zhuang and Mateu (2019) by not only allowing past events in the trigger, but also additional event types. Here, we explicitly consider the effect of follow-up visits by police, in addition to the self-exciting effect of reporting domestic abuse itself. Doing so requires us to consider additional event times and locations, those of police follow-ups. This results in an additional sequence of event times $\{t'_1, \dots, t'_k\}$ and of locations $\{s'_1, \dots, s'_k\}$,

where (t'_j, s'_j) give time and location of a follow-up event j . These different event types are distinguished by a sequence of length $n + k$ of labels M_j , where $M_j = 0$ if event j is a report of domestic abuse to police and $M_j = 1$ if event j is a follow-up.

Note that despite the introduction of additional events, we are still only modelling the intensity of domestic abuse reports (events for which $M_j = 0$). Allowing police follow-ups ($M_j = 1$) to induce additional reports of domestic abuse does not change the outcome event of interest. However, its introduction creates a notational challenge: The number of events to sum over differ between the background which only depends on the reports and the trigger which depends on outcome events (reports) and additional events (follow-ups). If we wanted to be absolutely precise in our notation, we would have to distinguish two sequences of event times: t^{reports} which holds the event times of our outcome event of interest and t^{all} which is the time-ordered sequence of *all* events in $t^{\text{reports}} \cup t^{\text{followups}}$. To then denote the total triggering probability mass on event i , we would need to write:

$$\sum_{\substack{j:t_j < t_i \\ t_j \in t^{\text{all}}}} \theta_{M_j} g(t_i - t_j) h(s_i - s_j).$$

In an attempt to avoid such cumbersome indexing, we abuse notation and assume that anytime we index over j , we are implicitly indexing over all events in t^{all} . Using this shortcut, we instead rewrite the last equation to:

$$\sum_{j:t_j < t_i} \theta_{M_j} g(t_i - t_j) h(s_i - s_j),$$

where the initial

$$g(t_i - t_j) = \frac{1}{(t_i - t_j)/24 + 1/24}, \quad t_i > t_j$$

$$h(s_i - s_j) = \frac{1}{1 + (s_i - s_j)^2}.$$

We measure distance in kilometres and time in days, such that $t_i = 1.5$ denotes the time of an event i that took place 1.5 days (= 36 hours) since the beginning of the study window. In contrast with e.g., Kalair, Connaughton, and Di Loro (2020), we do not enforce monotonicity of $g(t)$ and $h(s)$ since we might expect the social dynamics of reporting to be non-monotonic.

Together, we finally have the conditional intensity function:

$$\lambda(t, s) = m_0 \mu_{\text{trend}}(t) \mu_{\text{weekly}}(t) \mu_{\text{daily}}(t) \mu_{\text{area}}(s) + \sum_{j:t_j < t} \theta_{M_j} g(t - t_j) h(s - s_j). \quad (2.3)$$

2.4.2 Inference

Performing inference for this model is challenging: The estimation of the background component requires that one can distinguish events coming from the background and triggered events (Reinhart, 2018). This paper follows the inference procedure first proposed in Zhuang, Ogata, and Vere-Jones (2002) and applied in the context of crime in Zhuang and Mateu (2019). While full details are available in these references, the following section provides a brief overview over the inference procedure's key steps. We begin with stochastic declustering, which gives the answer to the question of why an iterative procedure allows us to obtain estimates for our model components. We then explain how the exact estimates for the model components are derived and then put forward our extension of the model.

Stochastic declustering

When the background component contains a non-parametric element estimated from data, we need to be able to separate out the events coming from the background to properly estimate it. Zhuang, Ogata, and Vere-Jones (2002) propose the following basic approach: With the observed realizations of the point process and the conditional intensity as defined in Equation (2.1), we can define two quantities of interest from this setup.

1. the probability that an event came from the background, rather than the trigger component,

$$\varphi_i = \text{P}(\text{event } i \text{ came from background}) = \frac{\mu(t_i, s_i)}{\lambda(t_i, s_i)} \quad (2.4)$$

2. the probability that an event was triggered by a past event

$$\rho_{ij} = \text{P}(\text{event } i \text{ was triggered by } j, j < i) = \frac{f(t_i - t_j, s_i - s_j)}{\lambda(t_i, s_i)}. \quad (2.5)$$

By the law of total probability, it is clear that

$$\varphi_i + \sum_{j=1}^{i-1} \rho_{ij} = 1. \quad (2.6)$$

One way to interpret Equation (2.6) is that all events $j = 1, \dots, i-1$ preceding event i added probability mass ρ_{ij} on the event i , which allows us to decompose event i into background and trigger.

We can now begin to address our problem: In order to estimate the background, we need to determine whether an event came from the background (φ_i), which in turn depends on f (to obtain λ). The procedure is therefore iterative, beginning with an initial guess for μ , f and λ and iterating until convergence.

A very naive first guess at an estimator for μ would be a histogram estimator. For a spatio-temporal point process with conditional intensity as in Equation (2.1), we can subdivide the spatial study area into K subdivisions S_k and assume that the background is piece-wise constant in each subdivision. A histogram estimator for subdivision k is then given by (compare Equation (31) in Zhuang, 2020):

$$\hat{\mu}_k = \frac{1}{\|S_k\|} \sum_i \hat{\varphi}_i \mathbb{I}((t_i, s_i) \in S_k),$$

where \mathbb{I} is the indicator function. While instructive, this estimator is always coarser than and therefore inferior to a kernel density estimator:

$$\hat{\mu}(s, t) = \sum_i \varphi_i Z(t - t_i; b_t) Z(s - s_i; b_i),$$

where

$$Z(x; b) = \frac{1}{\sqrt{2\pi}b} \exp\left(-\frac{x^2}{2b^2}\right)$$

is a Gaussian kernel with bandwidth b . b_t denotes the bandwidth for the temporal kernel and b_i is an event-specific bandwidth for the spatial kernel. This adaptive bandwidth accounts for the fact that a single bandwidth is often a poor choice with clustered point processes because it oversmooths some areas while being too noisy in other areas (Reinhart, 2018). Instead, b_i is set such that a spatial disk centred on event i with radius b_i contains n_p other events (Zhuang, Ogata, and Vere-Jones, 2002; Zhuang, 2011).

In much the same way, we can construct an estimator for f based on the triggering probabilities ρ_{ij} :

$$\hat{f}(t, s) = \sum_{i,j} \rho_{ij} Z(t - (t_i - t_j); b_g) Z(s - (s_i - s_j); b_h),$$

where b_g and b_h are the bandwidths for the temporal kernel and the spatial kernel, respectively. With these estimators in hand, we can in principle define an iterative procedure: With an initial guess for μ and f , we calculate φ and ρ . Then we update our estimates for μ and f , update φ and ρ until convergence.

Yet, this precise inference procedure does not actually work for the concrete model proposed: The iterative procedure can only recover a single temporal component applied to the entire period. However, in our specification of $\mu(t, s)$ in Equation (2.2), we specified multiple periodic components to account for periodicity. To obtain estimates for all background components, we need to modify our inference procedure.

Estimating periodic non-parametric background

To do that, we rely on the Georgii-Nguyen-Zessin formula (Georgii, 1976; Nguyen and Zessin, 1979) developed in the context of spatial and spatio-temporal point processes by Baddeley et al. (2005) and Zhuang (2006):

$$\mathbb{E} \left[\int_{[T_1, T_2] \times X} \gamma(t, s) dN(t \times s) \right] = \mathbb{E} \left[\int_{T_1}^{T_2} \int_X \gamma(t, s) \lambda(t, s) dt ds \right], \quad (2.7)$$

for a time interval $[T_1, T_2]$, area X and a non-negative function γ . Equation (2.7) is not trivial. Briefly, say we would like to know the value of the function γ over the entire space over which our point process is defined. Equation (2.7) states that we can evaluate the function γ over all observed points and the expectation of this (i.e., the left-hand side of Equation (2.7)) is equivalent to the expectation of that function over the entire space weighted by the intensity of the point process. This allows us to then rearrange terms and obtain the expectation of the function over the entire space.

Going back to our inference problem, recall that we are still trying to estimate the model components. For each of these components, we can construct a non-negative function w which is the component's contribution to the overall intensity.

As an example, take the trend component:

$$w^{\text{trend}}(t, s) = \frac{\mu_{\text{trend}}(t)\mu_{\text{area}}(s)}{\lambda(t, s)}. \quad (2.8)$$

Now we can substitute w for γ in Equation (2.7), by considering the time interval $[t - \Delta_t, t + \Delta_t]$, where Δ_t is a small positive number, and the whole of domain X to obtain:

$$\begin{aligned} \sum_i w^{\text{trend}}(t_i, s_i) \mathbb{I}(t_i \in [t - \Delta_t, t + \Delta_t]) &\approx \int_{T_1}^{T_2} \int_X w^{\text{trend}}(u, v) \lambda(u, v) \\ &\quad \mathbb{I}(u \in [t - \Delta_t, t + \Delta_t]) du dv \\ &= \int_{t - \Delta_t}^{t + \Delta_t} \mu_{\text{trend}}(u) du \int_X \mu_{\text{area}}(v) dv \\ &\propto \int_{t - \Delta_t}^{t + \Delta_t} \mu_{\text{trend}}(u) du \\ &\approx \mu_{\text{trend}}(t) 2\Delta_t. \end{aligned} \quad (2.9)$$

We can then rearrange the last expression to

$$\hat{\mu}_{\text{trend}}(t) \propto \sum_i \underbrace{\frac{\mu_{\text{trend}}(t_i)\mu_{\text{area}}(s_i)}{\lambda(t_i, s_i)}}_{:= w_i^{\text{trend}}} \mathbb{I}(t_i \in [t - \Delta_t, t + \Delta_t]). \quad (2.10)$$

Finally, we can smooth our estimates by using kernel density estimates which replace the indicator function in Equation (2.10) to obtain:

$$\hat{\mu}_{\text{trend}}(t) \propto \sum_i w_i^{\text{trend}} Z(t - t_i; b_{\text{trend}}) \quad (2.11)$$

In a similar fashion to Equation (2.8), we can define functions w and then estimators for all background components:

$$\hat{\mu}_{\text{daily}}(t) \propto \sum_i w_i^{\text{daily}} \sum_{k=0}^T Z(t - (t_i - \lfloor t_i \rfloor + k); b_{\text{daily}}) \quad (2.12)$$

$$\hat{\mu}_{\text{weekly}}(t) \propto \sum_i w_i^{\text{weekly}} \sum_{k=0}^{\lfloor T/7 \rfloor} Z(t - (t_i - 7\lfloor t_i/7 \rfloor + 7k); b_{\text{weekly}}) \quad (2.13)$$

$$\hat{\mu}_{\text{area}}(s) \propto \sum_i \varphi_i Z(s - s_i; b_{\text{area}}), \quad (2.14)$$

where $\lfloor x \rfloor$ denotes the largest integer smaller than or equal x . The periodicity

in μ_{daily} and μ_{weekly} comes from mapping the input event time t into the periodic domain. For the daily component, we simply subtract the day on which the event took place and are left with the time of day on which the day took place ($t_i - \lfloor t_i \rfloor$). For the weekly component, we subtract the week from the event time ($t_i - 7\lfloor t_i/7 \rfloor$).

For the triggering components, we obtain very similar expressions. As before, we can define a function

$$w^f(t, s, u, v) = \begin{cases} g(u - t)h(v - s)/\lambda(u, v) & \text{if } u > t, \\ 0 & \text{otherwise.} \end{cases}$$

We can then substitute w^f for γ into Equation (2.7) for a fixed t_i and s_i by considering the time interval $[t - \Delta_t, t + \Delta_t]$ to obtain

$$\begin{aligned} \sum_i w^f(t_j, s_j, t_i, s_i) \mathbb{I}(t_i - t_j \in [t - \Delta_t, t + \Delta_t]) &\approx \int_0^T \int_X w^f(t_j, s_j, u, v) \\ &\quad \mathbb{I}(u - t_j \in [t - \Delta_t, t + \Delta_t]) \lambda(u, v) du dv \\ &\approx \int_{t - \Delta_t}^{t + \Delta_t} g(u - t_j) du \int_X h(v - s_j) dv \\ &\propto g(t) \end{aligned}$$

To now obtain a more stable estimate for $g(t)$, we can additionally sum over the left-hand side for all j and obtain

$$g(t) \propto \sum_{i,j} w^f(t_j, s_j, t_i, s_i) \mathbb{I}(t_i - t_j \in [t - \Delta_t, t + \Delta_t]).$$

Now it is clear that $w^f(t_j, s_j, t_i, s_i)$ is just ρ_{ij} , which means that we can rewrite the last line to obtain an estimate for $g(t)$:

$$\hat{g}(t) \propto \sum_{i,j} \rho_{ij} \mathbb{I}(t_i - t_j \in [t - \Delta_t, t + \Delta_t]). \quad (2.15)$$

Similarly, we can obtain an estimate for $h(s)$:

$$\hat{h}(s) \propto \sum_{i,j} \rho_{ij} \mathbb{I}(s_i - s_j \in [s - \Delta_s, s + \Delta_s]), \quad (2.16)$$

where Δ_s is a small positive number. In addition, we apply a repetition correction which counts how often the triggering effect is observed at a specific time or space

distance (Zhuang and Mateu, 2019):

$$\hat{g}(t) \propto \frac{\sum_{i,j} \rho_{ij} Z(t - (t_i - t_j); b_g)}{\sum_j \mathbb{I}(t_j + t \leq T)} \quad (2.17)$$

$$\hat{h}(s) \propto \frac{\sum_{i,j} \rho_{ij} Z(s - (s_i - s_j); b_h)}{\sum_j \mathbb{I}((s_j + s) \in X)}. \quad (2.18)$$

Estimating the weighting terms and including additional event types

With the estimators for all background and triggering components, we can now turn to estimating the weighting terms m_0 and θ_M from Equation (2.3). The original paper by Zhuang and Mateu (2019) does not consider additional event types which is why we extend their model to accommodate the effect of follow-up visits by police.

The model defined in Equation (2.3) with complete parameter vector $\Theta = \{m_0, \theta_0, \theta_1\}$ has the following log-likelihood (Daley and Vere-Jones, 2003):

$$\ell(\Theta) = \sum_i \log \lambda(t_i, s_i) - \int_0^T \int_X \lambda(t, s) dt ds. \quad (2.19)$$

For the interested reader, we write out the full likelihood in Section 2.A but for brevity, we set the derivative of Equation (2.19) with respect to m_0 and θ_0 to zero (the derivative for θ_1 is analogous):

$$\begin{aligned} \partial \ell(\Theta) / \partial m_0 &= 0 \\ &= \sum_i \frac{\mu_{\text{trend}}(t_i) \mu_{\text{weekly}}(t_i) \mu_{\text{daily}}(t_i) \mu_{\text{area}}(s_i)}{\lambda(t_i, s_i)} \\ &\quad - \int_0^T \int_X \mu_{\text{trend}}(t) \mu_{\text{weekly}}(t) \mu_{\text{daily}}(t) \mu_{\text{area}}(s) dt ds \end{aligned} \quad (2.20)$$

and

$$\begin{aligned} \partial \ell(\Theta) / \partial \theta_0 &= 0 \\ &= \sum_i \frac{\sum_{j:t_j < t_i} \mathbb{I}(M_j = 0) g(t_i - t_j) h(s_i - s_j)}{\lambda(t_i, s_i)} \\ &\quad - \int_0^T \int_X \sum_{j:t_j < t} \mathbb{I}(M_j = 0) g(t - t_j) h(s - s_j) ds dt. \end{aligned} \quad (2.21)$$

Similar to Zhuang and Mateu (2019), this system of equations can be solved by basing the estimates for m_0 , θ_0 and θ_1 in inference round $(k + 1)$ on estimated

quantities from round (k) .

Kernel bandwidths and edge correction

Throughout the previous sections, we introduced several kernel bandwidths b_{daily} , b_{trend} , b_{weekly} , b_g and b_h . Typically, we would use cross-validation to choose values for those bandwidths. Given the computational complexity of the model however, this is not feasible. One alternative would be using a rule of thumb or other heuristic choice, but we can actually do better than that. Because the kernels governed by b_{daily} , b_{trend} and b_{weekly} are defined on a timeline, they are univariate. This means that the bandwidths for our Gaussian kernels retain an interesting interpretation: We can choose bandwidths with respect to the temporal range that we want the kernel to smooth over. For example, we set $b_{\text{daily}} = 1/24 \approx 0.04$ such that one standard deviation corresponds to 1 hour. That implies that 99.7% (= 3 standard deviations) of the contributions to our kernel density estimate come from events within 3 hours of the event. Similarly, we set $b_{\text{weekly}} = 1/3 \approx 0.33$ which corresponds to 99.7% of the contributions to the weekly kernel density estimate to come from events within 24 hours around our event. Lastly, we select $b_{\text{trend}} = 10$ which implies that 99.7% of the contributions to the trend kernel density estimate come from events within 30 days of our event.

As already discussed in Section 2.4.2, b_{area} is set to be adaptive so it does not require an explicit choice. However, it does depend on choice of n_p , the number of neighbours to the event. Zhuang (2011) propose setting n_p between 3 and 6, and our model uses $n_p = 5$.

Lastly, we have to choose b_g and b_h . Since the kernels are not Gaussian, the straightforward interpretation of the previous bandwidths does not work. Still, we set the bandwidth for the temporal distance between events to $b_g = 1$, corresponding to one day and $b_h = 0.2$, corresponding to 200m in spatial distance.

Because kernel density estimates are well-known to behave poorly around the edges, an edge correction is necessary. For the periodic and area kernels and the kernel smoothing $g(t)$ and $h(s)$, we use a truncated kernel which normalizes the kernel density estimator by its integral over the support (Hall and Turlach, 1999). For example, the estimator for μ_{daily} from Equation (2.12) is modified to:

$$\hat{\mu}_{\text{daily}}(t) \propto \sum_i w_i^{\text{daily}} \frac{\sum_{k=0}^T Z(t - t_i + \lfloor t_i \rfloor - k; b_{\text{daily}})}{\int_0^T Z(u - t_i; b_{\text{daily}}) du}. \quad (2.22)$$

This modification was not sufficient to ensure sensible edge behaviour for the trend kernel. That is because the support for the trend kernel is bounded between $[0, T]$. Instead, we apply an edge correction proposed by Schuster (1985) called boundary folding where the density “leaking” outside the support is mirrored or folded back onto the support. For a kernel with support in $[a, b]$, we correct the standard kernel density estimator $f_h(x) = \frac{1}{nh} \sum_i K(\frac{x-x_i}{h})$ to

$$f_h(x) = \frac{1}{nh} \sum_i K\left(\frac{x - 2a + x_i}{h}\right) + K\left(\frac{x - x_i}{h}\right) + K\left(\frac{x + 2b - x_i}{h}\right). \quad (2.23)$$

2.5 Results

With this inference procedure, we can now fit our model to the data. The model was implemented in R and is publicly available at https://github.com/laravomfell/reporting_spillovers, together with a file that generates synthetic data since the original data cannot be provided publicly.

Besides the full model in Equation (2.3), we also estimated a model without the periodic components in the background specification. However, the AIC of the full model was considerably smaller (20,392 compared to 21,735) so in what follows, we only consider the full model.

In Figures 2.2a through 2.2e we visualize the estimated intensities coming from the background components. As shown in Equation (2.2), these estimated intensities are multiplied together and then weighted by m_0 . We estimate $m_0 = 0.1689$.

The trend component in Figure 2.2a demonstrates that there is no dominant trend in domestic abuse reports over our study period since most of the normalized Kernel density lies between 0.9 and 1.1, i.e., close to the mean of 1. We observe an increase in domestic abuse reporting, however, in the summer months of July and August.

We document strong time of day and day of week effects. There are remarkably few calls between the hours of midnight and 4am with the estimated intensity increasing during the day. We observe two peaks of intensity, one between 12:00 and 13:00 and another one between 20:00 and 21:00 before calls drop off at night again. Looking at the weekly periodicity, we observe a strong weekend effect as calls begin to pick up from Friday onwards throughout the weekend. Together, the daily and weekly periodicity visualized in Figure 2.2d show that Friday evenings,

Saturday evenings and Sunday mornings are particularly high-intensity periods for reports of domestic abuse.

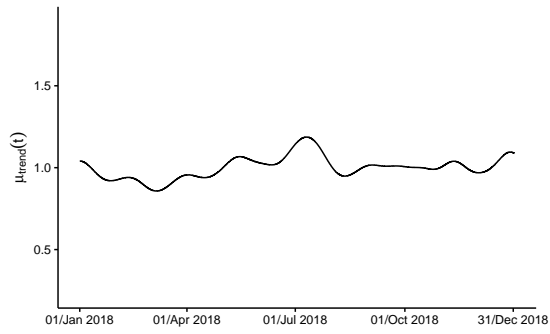
These findings mirror those found in other studies: Reports of domestic abuse are lowest during the week and strongly increase on the weekend with Sunday being the peak day (Rotton and Cohn, 2001; Brimicombe and Cafe, 2012).

Lastly, we find that there is marked variation in space. Specifically, we find that for some locations, the estimated spatial intensity of the background is particularly pronounced. This is clear from the small, dark spots in Figure 2.2e.

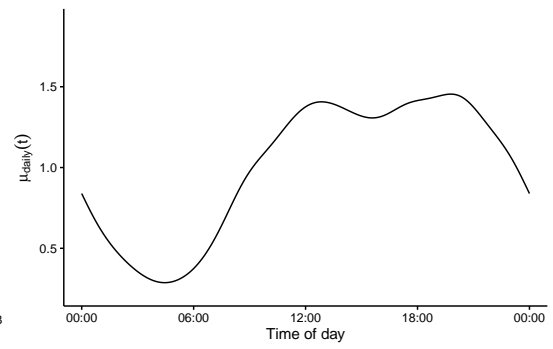
This raises a question of why these areas might see such high levels of domestic abuse reporting. A number of studies predict and confirm a relationship between levels of domestic abuse and deprivation (e.g., Gracia et al., 2015). In Figure 2.2f, we show the spatial background intensity at event locations against local deprivation. Deprivation is measured using the 2015 index of multiple deprivation, a composite index combining measures of deprivation from seven domains such as income, employment, education and health (Office for National Statistics, 2015). Overall the relationship between abuse reporting and deprivation is not straightforward: We see high levels of reporting in areas with high and low levels of deprivation. However, event locations with very high spatial background intensity (points in darker colours) are consistently in areas with high levels of deprivation. Regarding the triggering component, we visualize the estimated triggering functions $g(t)$ and $h(s)$ in Figure 2.3a and 2.3b. As shown in Equation (2.3), these functions are weighted by θ_M . Our estimated θ_M both for reports and for follow-ups are essentially zero: 1.52×10^{-9} for report-to-report and 2.19×10^{-8} for follow-up-to-report. Those numbers mean that one report of domestic abuse triggers, on average, 1.52×10^{-9} further reports. Similarly, one follow-up by police triggers, on average, 2.19×10^{-8} reports of domestic abuse. Together, the model implies that of the 6,084 initial reports of domestic abuse, $9.75 \times 10^{-7} \%$ reports were triggered by other events. In other words, there is very little evidence to support the notion of spillovers in domestic abuse reporting.

Furthermore, Figure 2.3a and 2.3b demonstrate that even before the weighting with θ , the estimated triggering functions imply very little triggering: The temporal range of triggering is very small to begin with and limited to the first 6 days after an event. Similarly, the spatial range of triggering is limited to an area of 400×400 m around an event. In summary, our model finds no evidence of spillovers in domestic abuse reporting.

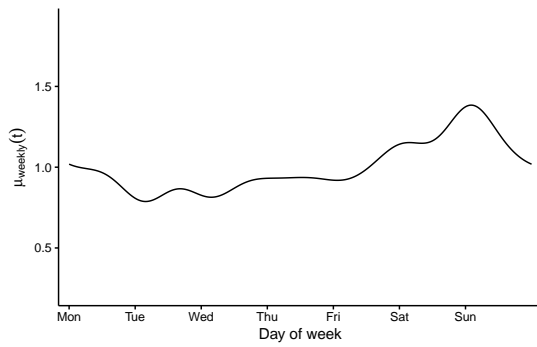
Since Hawkes processes are complex statistical objects with challenging inference procedures, it is important to validate the plausibility of model outputs.



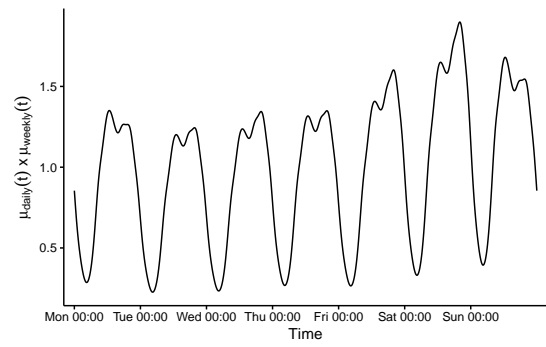
(a) Temporal trend



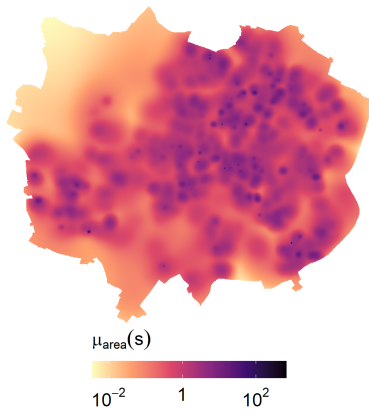
(b) Daily periodicity



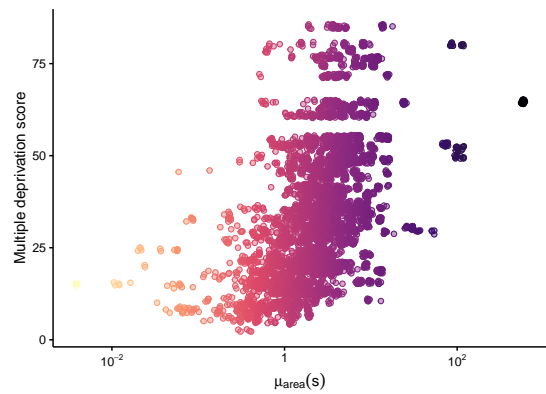
(c) Weekly periodicity



(d) Daily and weekly periodicity



(e) Spatial background intensity



(f) Spatial intensity at locations against deprivation

Figure 2.2: Estimated background components

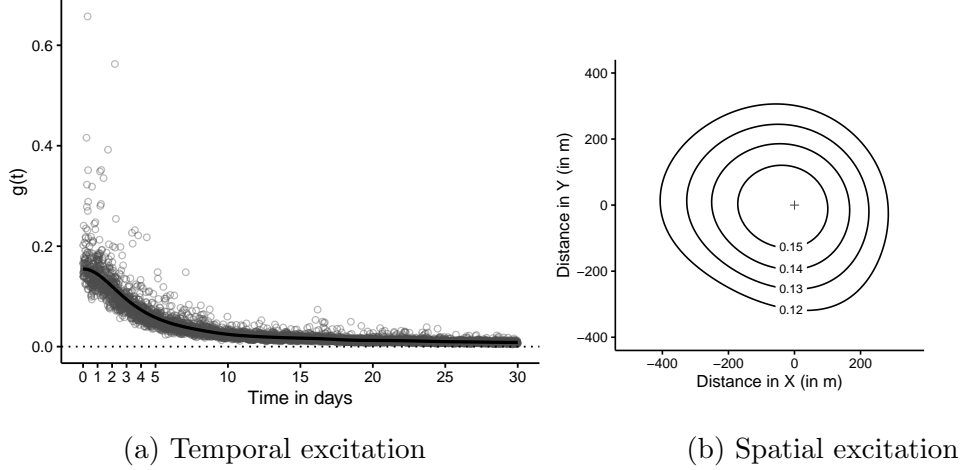


Figure 2.3: Estimated triggering components

For example, Reinhart and Greenhouse (2018) show that when the background component does not provide a good fit to the data, the triggering component is inflated. In other words, one typically over-estimates the triggering component. This is not the case for our model since our estimates of the triggering component are negligibly small. Still, we perform two additional plausibility checks to ensure that the quantities produced by our model are sensible.

First, we check for the plausibility of our triggering findings: We expect spillover effects to be stronger in areas where people are more aware of what is happening in their neighbours’ households. Since no data on this is available, we use the share of households living in non-detached houses in the neighbourhood as a proxy instead (Office for National Statistics, 2011b). People living in terraced houses or flats are much closer to any issues in their neighbours’ domestic life (Ivandic, Kirchmaier, and Linton, 2020). We define neighbourhood as 2011 Census Output areas with, on average, 300 usual residents in 150 households (Office for National Statistics, 2016). For each event, we then evaluate how many other reports it triggered according to our model by calculating $\theta_{M_j} \sum_i \rho_{ij}$ for each event j . This gives us a quantification of how much an event j increases the likelihood of further reports of domestic abuse around itself.

In Figure 2.4 we show the share of households living in non-detached houses in the area against the mean number of reports triggered. Indeed, we find weak evidence that follow-up events taking place in neighbourhoods where people live closer together exert more triggering pressure than events taking place in areas where houses are more spread out. While this effect is very small, it demonstrates that the quantities produced by our model are plausible.

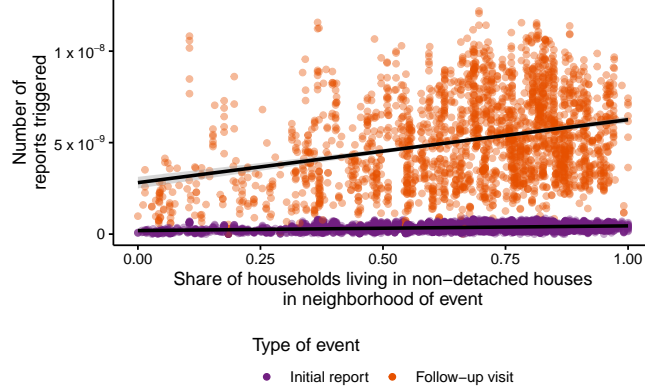


Figure 2.4: Share of households in neighbourhood living in non-detached houses against the mean number of reports triggered. Black lines show separate linear regression fits with 95% confidence intervals shown in gray.

As a second step, we check the model fit. A useful way of checking if the model is well-calibrated is a temporal residual plot. To do so, we calculate the following function of event times t_i

$$t_i \rightarrow \tau_i = \int_0^{t_i} \int_X \lambda(u, v) du dv, \quad (2.24)$$

such that τ_i is the expected number of events in the time interval $[0, t_i)$ or the cumulative number of events by event time t_i . This simple transformation makes use of the time-rescaling theorem: If the model is correct, the sequence of τ_i is a stationary Poisson process with unit rate (Ogata, 1988; Brown et al., 2002). Accordingly, a plot of the event index i against τ_i should form a 45 diagonal line. This property can be used to assess model fit: If $\hat{\lambda}$ is a good approximation of the true model, then its sequence of $\hat{\tau}_i$ will behave similarly to the sequence of theoretical τ_i (Schoenberg, 2002). For the theoretical τ_i we can construct confidence intervals for each τ_i by taking the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of a Beta distribution with parameters $(i + 1, n - i + 1)$ and then multiply by n (Zhuang and Mateu, 2019). We set α at the usual 0.05 level.

In Figure 2.5 we plot the deviation of $\hat{\tau}_i$ from the diagonal against the event index i to verify how far away from the diagonal our model deviates. Indeed, we find that $\hat{\tau}_i$ is within the 95% confidence bounds of the true model and that our model is therefore a reasonable approximation.

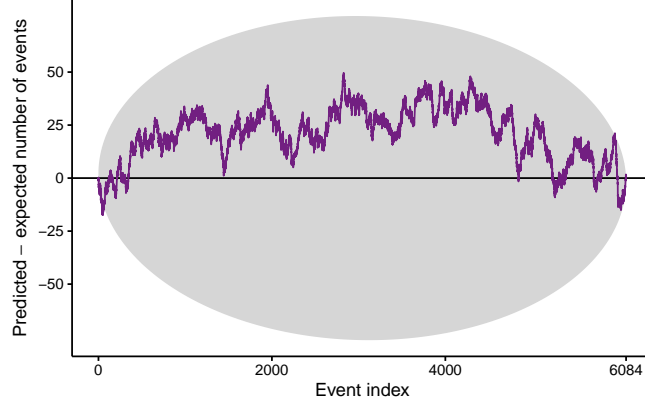


Figure 2.5: Deviation of the transformed time sequence (purple) from the theoretical (black) sequence with 95% confidence bands (grey).

2.6 Discussion

This paper studies if—like the criminal behaviour of offenders—the reporting of crime by victims exhibits triggering behaviour. We particularly investigate whether there are any spillover effects in the reporting of domestic abuse to police.

Analysing data from one year of calls for service concerning domestic abuse in a large English city, we find no convincing evidence for spillover effects. Spillover effects are limited to a very short time frame (within 6 days) and very short distances (400m around the event). These effects do not plausibly account for spillovers due to information sharing by victims in a social or neighbourhood network. We find some very weak evidence that events taking place in more densely populated neighbourhoods increase the likelihood of further reports of domestic abuse slightly more than events taking place in less dense neighbourhoods. Overall, reporting of domestic abuse by victims does not appear to exhibit any triggering behaviour since only 9.75×10^{-7} % of the reports in our sample are predicted to have been triggered.

The estimation of the background intensity of domestic abuse reporting shows that reporting follows highly periodic patterns. Calls for service of domestic abuse increase on the weekend and particularly in the evening. We also find that the reporting of domestic abuse is highly clustered and some locations in our study area see very high levels of reported domestic abuse.

A natural question arising from our work is if our estimates of the background intensity are also modelling the spatio-temporal intensity of domestic abuse itself. Certainly, in some instances the timing of a report of domestic abuse will coincide

with the timing of the domestic abuse itself. A range of models of domestic abuse, from ecological to feminist to economic models, make predictions about when violence is likely to break out in the abuse cycle (Heise, 1998; Bobonis, González-Brenes, and Castro, 2013; Lombard and McMillan, 2013; Leonard and Quigley, 2017). However, we cannot meaningfully separate the incidence of domestic abuse from the incidence of domestic abuse reporting, especially because many survivors of domestic abuse will have experienced multiple abuse incidents before alerting the police, if at all (SafeLives, 2015).

There is good reason to believe that making the decision to report domestic abuse to police would also influence others suffering from abuse in their reporting decision. However, our study documents that none of these hypotheses are confirmed in the reporting of domestic abuse.

There are a few reasons why we did not find an effect: The functional form of the triggering component of the Hawkes process might not be able to accommodate the shape of spillover effects. In their study of the effect of the #MeToo movement, Levy and Mattsson (2020) find that the effect is largest on crimes reported a month after they took place. Therefore, it is possible that assuming temporal spillovers in such a limited time frame is an ill fit to the nature of spillovers in reporting.

It may also be that expecting reporting to police to affect further domestic abuse reporting is overly optimistic. Studies consistently show that victims have quite heterogeneous preferences and justice goals in mind when they approach formal institutions (Rajah, Frye, and Haviland, 2006; Kelly, Sharp-Jeffs, and Klein, 2014). Evidence has shown that survivors of domestic abuse are most satisfied with approaches that provide them with options (Ellsberg, Arango, et al., 2015). Police officers without special training may not be sufficiently attuned to respond to victims/survivors' agency. This is actually partially reflected in the studies in Davis, Weisburd, and Taylor (2008)' meta analysis: The programmes where only a police officer (i.e., not a social worker and a police officer together) visited the household did not have a significant effect of the reporting of future violence (e.g., Davis, Weisburd, and Hamilton, 2010; Pate, Hamilton, and Annan, 1992). If such visits already do not encourage victims in the treated households to turn to police again, the likelihood that such visits would have any effect outside those households is low.

Lastly and perhaps most important, it is worth examining the differences between crimes that exhibit strong triggering behaviours and domestic abuse. Clearly, the offender side is different: Unlike burglars, perpetrators of domestic

abuse do not choose the geographic location of their crimes. This was an explicit reason for choosing domestic abuse for our investigation. But the nature of the crime is also different: Burglaries, homicides and shootings are all discrete events with a distinct time and place of offence. This gives crime victims a concrete event to report. In contrast, domestic abuse consists of both discrete events such as assaults but more so of patterns of abusive behaviour. With this combination, domestic abuse cannot be thought of as a sequence of individual criminal offences (Hawkins and Laxton, 2014). The discretisation of a latent, on-going phenomenon such as domestic abuse into reports may not capture if and when victims' share information about domestic abuse and the police response to it.

Appendix 2.A Likelihood

We did not write out the full log likelihood in Equation (2.19). We do this here, but for simplicity, we simplify the background to be denoted by $m_0\mu(t, s)$ and the trigger by $\theta_M f(t, s)$. Altogether, our model in Equation (2.3) with parameter vector $\Theta = \{m_0, \theta_0, \theta_1\}$ has the following log likelihood:

$$\begin{aligned}
\ell(\Theta) &= \sum_i \log \lambda(t_i, s_i) - \int_0^T \int_X \lambda(t, s) dt ds \\
&= \sum_i \log \left(m_0 \mu(t_i, s_i) + \sum_{j:t_j < t_i} \theta_{M_j} f(t_i - t_j, s_i - s_j) \right) \\
&\quad - \int_0^T \int_X \left(m_0 \mu(t, s) + \sum_{j:t_j < t} \theta_{M_j} f(t - t_j, s - s_j) \right) dt ds \\
&= \sum_i \log(m_0 \mu(t_i, s_i)) + \sum_i \log \left(\sum_{j:t_j < t_i} \mathbb{I}(M_j = 0) \theta_0 f(t_i - t_j, s_i - s_j) \right) \\
&\quad + \sum_i \log \left(\sum_{j:t_j < t_i} \mathbb{I}(M_j = 1) \theta_1 f(t_i - t_j, s_i - s_j) \right) - \int_0^T \int_X m_0 \mu(t, s) dt ds \\
&\quad - \int_0^T \int_X \sum_{j:t_j < t} \mathbb{I}(M_j = 0) \theta_0 f(t - t_j, s - s_j) dt ds \\
&\quad - \int_0^T \int_X \sum_{j:t_j < t} \mathbb{I}(M_j = 1) \theta_1 f(t - t_j, s - s_j) dt ds.
\end{aligned}$$

We can now take the derivative of Equation (2.25) with respect to m_0 , using the chain rule:

$$\begin{aligned}\partial\ell(\Theta)/\partial m_0 &= 0 \\ &= \sum_i \frac{\mu_{\text{trend}}(t_i)\mu_{\text{weekly}}(t_i)\mu_{\text{daily}}(t_i)\mu_{\text{area}}(s_i)}{\lambda(t_i, s_i)} \\ &\quad - \int_0^T \int_X \mu_{\text{trend}}(t)\mu_{\text{weekly}}(t)\mu_{\text{daily}}(t)\mu_{\text{area}}(s)dt ds. \quad (2.25)\end{aligned}$$

Again, using the chain rule we can now take the derivative of Equation (2.25) with respect to θ_0 (the derivative with respect to θ_1 is analogous) and obtain:

$$\begin{aligned}\partial\ell(\Theta)/\partial\theta_0 &= 0 \\ &= \sum_i \frac{\sum_{j:t_j < t_i} \mathbb{I}(M_j = 0)g(t_i - t_j)h(s_i - s_j)}{\lambda(t_i, s_i)} \\ &\quad - \int_0^T \int_X \sum_{j:t_j < t} \mathbb{I}(M_j = 0)g(t - t_j)h(s - s_j)ds dt. \quad (2.26)\end{aligned}$$

Appendix 2.B Inference algorithm

We can write out the inference procedure for the model explained in Section 2.4.2 in algorithmic form. The procedure consists of two main steps: The initialisation and the inference loop.

In the initialisation stage, we obtain initial values for the daily, weekly, trend, area and triggering components. We then use those to calculate the entire background component $\mu(t, s)$. We need to do this calculation twice: Once to obtain $\mu(t_i, s_i)$, that is the background value at all events i and once more to obtain $\int \mu(t, s)dt ds$, that is the background integrated over the entire study area. We then repeat this step to obtain the trigger at all events i $f(t_i, s_i)$ and integrated over the study area $\int f(t, s)dt ds$.

With those quantities in hand, we then update m_0 and θ_M from some initial guesses and then calculate the intensity λ , again at all events i and integrated over the study area.

We then enter the inference loop where essentially the procedure repeats: We obtain updated values for the daily, weekly, trend, area and triggering components; we calculate the background and triggering components at the events and integrated over the study area. We update m_0 and θ_M , and calculate λ at all

events i and integrated over the study area. When m_0 and θ converge, we break the inference loop.

More formally, we write:

Algorithm 1 Inference algorithm

Input: n_p , b_{daily} , b_{weekly} , b_{trend} , b_{area} , b_g , b_h , m_0 and θ_M

Initialisation

Initialise components μ_{daily} , μ_{weekly} , μ_{trend} , μ_{area} , $g(t)$, $h(s)$,

Calculate background $\mu(s_i, t_i)$ and $\int_0^T \int_X \mu(s, t) dt ds$

Calculate trigger $g(t - t_i)h(s - s_i)$ and $\sum_i \int_{t_i}^T \int_X g(t - t_i)h(s - s_i) dt ds$

Update m_0 and θ_M

Calculate intensity $\lambda(t_i, s_i)$ and $\int_0^T \int_X \lambda(t, s) dt ds$

while not convergence **do**

 Update components μ_{daily} , μ_{weekly} , μ_{trend} , μ_{area} , $g(t)$, $h(s)$

 Calculate background $\mu(s_i, t_i)$ and $\int_0^T \int_X \mu(s, t) dt ds$

 Calculate trigger $g(t - t_i)h(s - s_i)$ and $\sum_i \int_{t_i}^T \int_X g(t - t_i)h(s - s_i) dt ds$

 Update m_0 and θ_M

 Calculate intensity $\lambda(t_i, s_i)$ and $\int_0^T \int_X \lambda(t, s) dt ds$

 Check convergence of m_0 and θ_M

Chapter 3

Robust Bayesian Inference for Discrete Outcomes with the Total Variation Distance

3.1 Introduction

Discrete outcomes such as counts or classification labels pose significant modelling challenges because standard inference is vulnerable to over-weighting subtle data features such as boundary or censoring effects, zero-inflation, as well as issues such as outliers, inliers and corrupted data. This modelling difficulty is relevant to a wide range of applied fields, including crime where data are notoriously noisy.

Building on a growing literature on robustness in Bayesian models, the current paper provides a generic strategy for robustness in a discrete setting: In the absence of more detailed knowledge on the nature of misspecification which could then be explicitly modelled, we ensure robustness implicitly via the learning criterion. As outlined in Jewson, Smith, and Holmes (2018), a particularly appealing way of doing so is by way of a generalised Bayesian approach based on robust discrepancies or divergences (Bissiri, Holmes, and Walker, 2016). In the context of continuous data, this idea was first pioneered using α -divergences (Hooker and Vidyashankar, 2014) and has since been extended to β - and γ -divergences (Ghosh and Basu, 2016; Futami, Sato, and Sugiyama, 2018; Knoblauch, Jewson, and Damoulas, 2018; Knoblauch, Jewson, and Damoulas, 2019; Manousakas and Mascolo, 2020) as well the Maximum Mean Discrepancy (Chérif-Abdellatif and Alquier, 2020b). For various reasons, all these approaches are somewhat unattractive for statistical machine learning problems with discrete-valued data:

The approach of Hooker and Vidyashankar (2014) relies on a computationally inefficient kernel density estimate, β - and γ -divergences have no easily computable form outside the exponential family, and the Maximum Mean Discrepancy relies on kernels in a way that makes it less obvious how to work with it in discrete settings.

In light of these limitations and as illustrated in Figure 3.1, we propose a generalised Bayesian inference method for discrete-valued outcomes based on the Total Variation Distance (TVD). We make three contributions:

1. We explore the theoretical properties of our estimator for the TVD and find that it satisfies exponential concentration inequalities (Propositions 2 and 3), converges almost surely (Corollary 1) and retains the robustness properties of the true TVD (Corollary 2). Further, the estimator’s minimiser is strongly consistent (Proposition 4).
2. We adapt Bayesian Nonparametric Learning (NPL) as popularized by Lyddon, Walker, and Holmes (2018) and Fong, Lyddon, and Holmes (2019) to our setting. As the resulting algorithm is computationally equivalent to the Bayesian Bootstrap, inference has low computational complexity and is embarrassingly parallel.
3. We apply the new inference scheme to a range of simulated and real world data sets. As expected, the TVD yields superior performance under misspecification. Even in the absence of misspecification, we match performance of standard inference using the Kullback-Leibler Divergence (KLD).

In Section 3.2, we briefly recap divergence-based generalisations of Bayesian inference. Next, Section 3.3 motivates the use of the TVD within this framework. As the TVD needs to be estimated, we prove a number of properties exhibited by our estimator in Section 3.4. Section 3.5 explains how to embed our estimator into Bayesian Nonparametric Learning. We then apply the resulting algorithm to a range of simulated and real-world data sets, including a challenging crime inference example, and discuss the results in Section 3.6.

We conclude that the method constitutes a reliable and generic robustness strategy for inference in discrete outcome models.

To aid the reader through the statistical quantities used throughout this chapter, we provide an overview of the objects in Table 3.1.

Quantity	Definition	Explanation
\mathcal{X}		support of x
$p(x, y)$		true joint distribution
$p^x(x)$		true marginal distribution of x
p^y		true marginal distribution of y
$p^{y x}(y x)$		true conditional distribution of $y x$
$f_\theta(y x)$		parametric model for $y x$ parametrised by some $\theta \in \Theta$
$\hat{p}_n(x, y) = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}(x, y)$		empirical measure of the sample
$\hat{p}_n^x(x) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}(x)$		empirical marginal distribution of x
$\hat{p}_n^{y x}(y x) = \frac{\hat{p}_n(y, x)}{\hat{p}_n^x(x)}$		empirical conditional distribution $y x$
$p_\theta(x, y) = f_\theta(y x)p^x(x)$		‘hybrid’ distribution using the model conditional but the true marginal of x
$p_{\theta, n}(x, y) = f_\theta(y x)p_n^x(x)$		‘hybrid’ distribution using the model conditional but the empirical marginal of x
$\hat{p}_{n, h_n}^x(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{h_n^d} K\left(\frac{x_i - x}{h_n}\right)$		KDE for $p(x)$ over \mathcal{X}
$\hat{p}_{n, h_n}(x, y) = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}(y) \hat{p}_{n, h_n}^x(x)$		estimator for $p(x, y)$ where we use a KDE estimator over \mathcal{X}
$\hat{p}_{n, h_n}^{y x}(y x) = \frac{\hat{p}_{n, h_n}(y, x)}{\hat{p}_{n, h_n}^x(x)}$		estimator for $p(y x)$ where we use a KDE estimator over \mathcal{X}
$\hat{p}_{\theta, n, h_n}(x, y) = f_\theta(y x) \hat{p}_{n, h_n}^x(x)$		‘hybrid’ between the model and a KDE estimator of conditional
$\hat{p}_{n, h_n}^y = \int \hat{p}_{n, h_n}(x, y) dx = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}(y)$		marginalizing out x in $\hat{p}_{n, h_n}(x, y)$ to obtain marginal over y
$\hat{p}_{n, h_n}^{x y}(x y) = \frac{\hat{p}_{n, h_n}(y, x)}{\hat{p}_{n, h_n}^y(y)}$		estimator for $p(x y)$ where we use a KDE estimator over \mathcal{X}

Table 3.1: Overview of statistical quantities used in Chapter 3.

3.2 Divergences & Inference

It is well-known that both Maximum Likelihood estimation and conventional Bayesian Inference minimise the Kullback-Leibler divergence (KLD) between the empirical density \hat{p}_n and a model family $\{f_\theta : \theta \in \Theta\}$.

The Maximum Likelihood objective is given by

$$\begin{aligned} L_n(\theta) &= \frac{1}{n} \sum_{i=1}^n \log f_\theta(y_i|x_i) \\ &= \mathbb{E}_{\hat{p}_n^x} \left[\underbrace{\mathbb{E}_{\hat{p}_n^{y|x}} \left[\log \left(\frac{f_\theta(y|x)}{\hat{p}_n(y|x)} \right) \right]}_{=-\text{KLD}(\hat{p}_n^{y|x} \| f_\theta)} \right] - H(\hat{p}_n), \end{aligned} \quad (3.1)$$

where H denotes the Shannon-entropy, and $\hat{p}_n(y, x) = n^{-1} \sum_{i=1}^n \delta_{(x_i, y_i)}(x, y)$ defines the joint empirical distributions of a sample $\{y_i, x_i\}_{i=1}^n$. $\hat{p}_n^x(x) = n^{-1} \sum_{i=1}^n \delta_{x_i}(x)$ gives the marginal and $\hat{p}_n^{y|x}(y|x) = \hat{p}_n(y, x) / \hat{p}_n^x(x)$ the conditional empirical distributions of the sample. As $H(\hat{p}_n)$ does not depend on θ , Equation (3.1) shows that maximizing L_n over θ amounts to minimising $\mathbb{E}_{\hat{p}_n^x}[\text{KLD}(\hat{p}_n^{y|x} \| f_\theta)]$ over θ . Because the KLD is not robust to outliers and misspecification, this observation has inspired numerous alternative disparity-based techniques. While these methods were initially focused on statistical testing procedures (Wolfowitz, 1957), they were quickly extended in order to derive robust estimators (e.g., Beran, 1977; Yatracos, 1985; Simpson, 1987; Basu et al., 1998; Hyvärinen, 2005; Briol et al.,

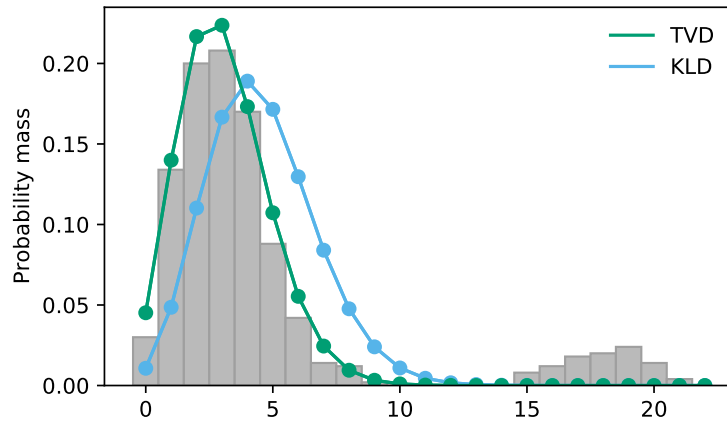


Figure 3.1: Discrete count data with some outliers (grey histogram) are modelled with a Poisson distribution. Inferring the model with a standard approach amounts to minimising the **KLD** between model and data. For this case, the outliers have a disproportionate impact. Minimising the **TVD** instead is robust to such contamination.

2019; Barp et al., 2019; Chérif-Abdellatif and Alquier, 2020a).

This extends to Bayesian inference: Defining $\pi(\theta)$ as the prior distribution, one can rewrite the Bayesian posterior given by $q_n(\theta) \propto \prod_{i=1}^n f_\theta(y_i|x_i)\pi(\theta)$ in similar fashion. Specifically, q_n solves a well-known variational problem: For $\mathcal{P}(\Theta)$ denoting the set of all probability measures on Θ ,

$$q_n(\theta) = \operatorname{argmin}_{q \in \mathcal{P}(\Theta)} \left\{ \mathbb{E}_q[\text{KLD}(\hat{p}_n \| f_\theta)] + \frac{1}{n} \text{KLD}(q \| \pi) \right\}.$$

Notice that this Bayesian objective is equivalent to a prior-regularized version of its Frequentist counterpart. This perspective on Bayesian inference is particularly popular within the PAC-Bayesian and Variational Inference communities (see e.g.,

Germain et al., 2016; Knoblauch, Jewson, and Damoulas, 2019; Guedj, 2019). In fact, the similarities with the Frequentist version of the objective ensures that their solutions coincide as $n \rightarrow \infty$ (see e.g., Ghosal, 1996), even if the model is misspecified (e.g., Ghosal and Van der Vaart, 2007), even if the discrepancy term assessing the model fit is no longer the KLD (e.g., Ghosh and Basu, 2016; Miller, 2019), and even if the prior regularization term is no longer the KLD (Knoblauch, 2019).

It is important to note this asymptotic equivalence: It implies that model misspecification issues plaguing Frequentist estimators for θ will carry over into Bayesian inference on θ if n is large enough.

3.3 Motivation

To address such robustness concerns for discrete data, the current paper studies generalised Bayesian inference based on the Total Variation Distance (TVD). Letting $p(y, x)$, $p^{y|x}(y|x)$ and $p^x(x)$ denote the distributions of the true joint, conditional and marginal data generating mechanism and $p_\theta(x, y) = f_\theta(y|x)p^x(x)$, this means that we want to produce high posterior density in the regions of Θ where

$$\begin{aligned} \text{TVD}(p, p_\theta) &= \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} |p^{y|x}(y|x) - f_\theta(y|x)| dp^x(x) \\ &= \mathbb{E}_{p^x} [\text{TVD}(p^{y|x}(\cdot|x), f_\theta(\cdot|x))] \end{aligned} \quad (3.2)$$

takes relatively small values.

While the list of potential robust discrepancy measures is long, the TVD is perhaps uniquely appealing. First, the TVD’s very definition shows that it will seek values of θ producing well-calibrated probability models: As the TVD is the average absolute difference between the candidate probability model f_θ and the true data generating distribution, it assigns the highest posterior density to values of θ that best describe how p allocates its probability mass. This feature is not only intuitively attractive, but particularly suitable for choosing amongst probability measures on discrete spaces. Second, the TVD does not depend on any additional hyperparameters. Accordingly, we require no cross-validation or tuning strategy before inference is performed. This is in stark opposition to a majority of alternatives: α -, β - and γ -divergences are named after their hyperparameters while Minimum Stein Discrepancies (Barp et al., 2019) and Maximum Mean Discrepancies (Briol et al., 2019) depend on the choice of a kernel. Third, and as we shall demonstrate next, the TVD has universal robustness guarantees that are far stronger than those of most alternatives.

Suppose the data-generating process is given by

$$c(y, x) = (1 - \varepsilon) \cdot f_{\underline{\theta}}(y|x)p^x(x) + \varepsilon \cdot q(y, x), \quad (3.3)$$

where $\varepsilon \in (0, 1)$ is the size of the contamination given by the distribution q , $f_{\underline{\theta}} \in \{f_\theta : \theta \in \Theta\}$ adequately describes the remaining data, and p^x is the uncontaminated marginal distribution on \mathcal{X} . Unlike with other discrepancy measures, any adverse effect q has on inferring θ via the TVD is bounded.

Proposition 1. *For $p_\theta(y, x) = f_\theta(y|x)p^x(x)$,*

$$|\text{TVD}(c, p_\theta) - \text{TVD}(p_{\underline{\theta}}, p_\theta)| \leq 2\varepsilon.$$

This result makes intuitive sense if one recalls that the TVD selects for values of θ that correctly match the probability mass of the data-generating distribution. Since the probability mass of contamination relative to the family $\{f_\theta : \theta \in \Theta\}$ is exactly ε , it logically follows that the TVD should be off by a factor of order at most ε . More striking still: The degree by which the contaminant q is different from the unpolluted component has no impact, a property that makes the TVD markedly different from the KLD (see e.g., Figure 3.1).

3.4 Estimating the TVD

These strong robustness properties make the TVD a uniquely appealing discrepancy measure. Unfortunately, we do not know the true data generating process p , which is needed to compute $\text{TVD}(p, p_\theta)$ exactly. Accordingly, we instead need to estimate $\text{TVD}(p, p_\theta)$. While this has inspired theoretically convincing prior work on estimating $\text{TVD}(p, p_\theta)$ (e.g., Yatracos, 1985; Devroye and Lugosi, 2012), the resulting estimators are typically both practically and computationally infeasible. For instance, the estimator introduced by Yatracos (1985) not only requires the parameter space to be totally bounded, but also discretised.

To ensure practicality, the current paper uses an estimator that is theoretically inferior, but computationally superior by orders of magnitude. With $\hat{p}_n(y, x)$, $\hat{p}_n^{y|x}(y|x)$ and $\hat{p}_n^x(x)$ as before and for $\hat{p}_{\theta,n}(y, x) = f_\theta(y|x)\hat{p}_n^x(x)$, we use

$$\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) = \mathbb{E}_{\hat{p}_n^x} [\text{TVD}(\hat{p}_n^{y|x}(\cdot|x), f_\theta(\cdot|x))] , \quad (3.4)$$

which is similar to the KLD estimator in Equation (3.1). Throughout, the *true* domain of the observations y_i is given by \mathcal{Y}_* , which may be a subset of the *model's* domain \mathcal{Y} . This is natural under model misspecification: For example, one may fit a Poisson regression model ($\mathcal{Y} = \mathbb{N}$) to the number of rainfall days in a year ($\mathcal{Y}_* = \{1, \dots, 365\}$) for n years.

While the TVD has many desirable theoretical properties, it does have one decisive drawback relative to the KLD: While the KLD loss of Equation (3.1) is *linear* in the log likelihood functions, the TVD loss of Equation (3.4) is non-linear. An immediate consequence is that performance guarantees of the TVD estimator—such as convergence properties or finite-sample bounds—are much harder to derive.

The results of this section show that in spite of these complications, our estimator is generally very reliable. A minor complication arises when the covariates $\{x_i\}_{i=1}^n$ are continuously-valued: In this case, we cannot study the estimator directly. Instead, we derive results for a surrogate estimator that relies on kernel density estimation, but can be made arbitrarily close to our naive estimator.

We find that under mild conditions, our estimator for the TVD satisfies exponential concentration inequalities. This immediately allows us to conclude that (i) it converges to its target almost surely and that (ii) the robustness property of Proposition 1 applies to the estimated objects, too. With additional labour, one can also show that the minimisers of $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$ are strongly consistent

for the minimiser of $\text{TVD}(p, p_\theta)$. Results with proof sketches are derived in full detail in Appendix 3.A.

3.4.1 Exponential Concentration Inequalities

As the arguments and rates are slightly different, we give separate results for discrete-valued and continuous-valued covariates.

Proposition 2. *If $\{x_i, y_i\}_{i=1}^n$ are both discrete-valued and sampled i.i.d. from a probability distribution \mathbb{P} such that $y_i \in \mathcal{Y}$, $x_i \in \mathcal{X}$ and $|\mathcal{Y}| = K_y$, $|\mathcal{X}| = K_x$ for some $K_x, K_y \in \mathbb{N}$, then it holds that pointwise for any $\theta \in \Theta$ and for any $\varepsilon > 0$ and with probability at least $1 - \delta_n$,*

$$|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| < \varepsilon$$

where $\delta_n = (2^{K_y + K_x + 1} - 2^2)e^{-n\varepsilon^2/2}$.

Proof sketch. One can use the triangle inequality and the fact that $|x + y| \geq ||x| - |y||$ to show that

$$|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| \leq \text{TVD}(\hat{p}_n, p) + \text{TVD}(p_\theta, \hat{p}_{\theta,n}).$$

For either of these two expressions, exponential concentration inequalities apply (e.g., Weissman et al., 2003) so that one can use a union bound to show that the result holds. \square

Similar arguments can be used for continuous covariates if one studies the surrogate estimator

$$\text{TVD}(\hat{p}_{n,h_n}, \hat{p}_{\theta,n,h_n}) \approx \text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$$

where for a suitable kernel K and bandwidth h_n , the kernel density smoothed estimator \hat{p}_{n,h_n} of p is

$$\hat{p}_{n,h_n}(x, y) = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}(y) \cdot \frac{1}{h_n^d} K\left(\frac{x_i - x}{h_n}\right).$$

Proposition 3. *Suppose $\{y_i\}_{i=1}^n$ are discrete with $|\mathcal{Y}| = K_y$ while $\{x_i\}_{i=1}^n$ are continuous with $\mathcal{X} \subseteq \mathbb{R}^d$ and that $\{x_i, y_i\}_{i=1}^n$ are sampled i.i.d. from a probability distribution \mathbb{P} . Assume that the true marginal distribution \mathbb{P}_x of x admits a*

density p^x that is absolutely continuous with respect to the Lebesgue measure. Further, suppose that $\int_{\mathcal{X}^2} K(x, x') dx dx' = 1$ and $h_n \rightarrow 0$ while $n \cdot h_n \rightarrow \infty$ as $n \rightarrow \infty$. Then it holds for n large enough, any $\theta \in \Theta$, any $\varepsilon > 0$ and with probability at least $1 - \delta_n$,

$$|\text{TVD}(\hat{p}_{n, h_n}, \hat{p}_{\theta, n, h_n}) - \text{TVD}(p, p_\theta)| < \varepsilon$$

where $\hat{p}_{\theta, n, h_n}(y, x) = p_\theta(y|x) \hat{p}_{n, h_n}^x(x)$. Here, for a constant r depending only on K_y , and for $n_y = \sum_{i=1}^n \delta_y(y_i)$ denoting the number of samples for which $y_i = y$, we have $\delta_n = \exp\{-\min_{y \in \mathcal{Y}} n_y \cdot \varepsilon^2 \cdot r\}$.

Proof sketch. The proof proceeds along similar lines as in the discrete case, but is complicated by the fact that for $q(x, y) = \hat{p}_{n, h_n}^{x|y}(x|y)p^y(y)$, one additionally upper bounds

$$\text{TVD}(\hat{p}_{n, h_n}, p) \leq \text{TVD}(\hat{p}_{n, h_n}, q) + \text{TVD}(q, p),$$

which in turn we can further upper bound by

$$\leq \text{TVD}(\hat{p}_{n, h_n}^y, p^y) + \sup_{y \in \mathcal{Y}} \text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y)).$$

The first term vanishes exponentially fast by the same arguments deployed for the discrete case. The second term requires a conditionalisation argument together with an upper bound which introduces $\min_{y \in \mathcal{Y}} n_y$ into the bound. Lastly, one uses a union bound argument to put everything together. \square

While the last result does not apply to the actual estimator of interest in Equation (3.4), it does apply to its kernel-density based surrogate. In spite of this, all experiments in the current paper use Equation (3.4) rather than its surrogate. Why do we insist on using an estimate of likely inferior theoretical quality?

The answer is threefold: Firstly, though a TVD estimate based on kernel density estimation is theoretically appealing, it would add two orders of magnitude to the computational complexity of our algorithm. This renders the kernel density surrogate practically infeasible for most situations, a feature present in pre-existing estimators for the TVD (e.g., Yatracos, 1985; Devroye and Lugosi, 2012). Secondly, it is reasonable to expect the behaviour of the naive estimates to be fairly similar to that based on kernel density estimates. In fact, one can

make \hat{p}_{n,h_n}^x arbitrarily close to \hat{p}_n^x for small enough h_n . For instance, Proposition 3 holds for a uniform kernel and $h_n = \eta \cdot n^{-1/5}$, which (even if $n = 1$) can be made arbitrarily close to the empirical measure for $\eta \rightarrow 0$. Thirdly and most persuasively, our empirical results demonstrate that the naive estimator of Equation (3.4) performs convincingly, even if $\mathcal{X} \subseteq \mathbb{R}^d$.

3.4.2 Almost sure convergence

The exponential concentration inequalities derived in Propositions 2 and 3 are attractive for many reasons. Most importantly, they provide computable finite-sample guarantees. Further, they also imply almost sure convergence by the Borel-Cantelli Lemma.

Corollary 1. *Under the conditions of Proposition 2, $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) \xrightarrow{a.s.} \text{TVD}(p, p_\theta)$ as $n \rightarrow \infty$. Under the conditions of Proposition 3, $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n,h_n}) \xrightarrow{a.s.} \text{TVD}(p, p_\theta)$ as $\min_{y \in \mathcal{Y}} n_y \rightarrow \infty$.*

3.4.3 Provable robustness

Proposition 1 demonstrated that the *true* TVD is robust. Further, Propositions 2 and 3 showed that the estimated TVD rapidly converges to the truth. Intuitively then, it stands to reason that the *estimated* TVD is also robust. The following result confirms this.

Corollary 2. *Pick any $\eta > 0$. Under the conditions of Proposition 2 and with \hat{p}_n an empirical measure constructed from c as in Equation (3.3), there is N such that for all $n \geq N$,*

$$|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p_{\underline{\theta}}, p_\theta)| \leq 2\varepsilon + \eta$$

holds with probability one. Similarly, under the conditions of Proposition 3, it holds that there exists N such that for all n for which $\min_{y \in \mathcal{Y}} n_y \geq N$,

$$|\text{TVD}(\hat{p}_{n,h_n}, \hat{p}_{\theta,n}) - \text{TVD}(p_{\underline{\theta}}, p_\theta)| \leq 2\varepsilon + \eta$$

holds with probability one.

3.4.4 Consistency

In spite of being firmly rooted within the Bayesian paradigm, the inference method we will present in the next section relies on computing a perturbed form of the minimisers $\theta_n = \operatorname{argmin}_{\theta \in \Theta} \operatorname{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$. Thus, we are interested in the convergence properties of θ_n . While the convergence properties of similar estimators have been studied before (Yatracos, 1985), the analysis we employ is drastically different. This is because unlike the estimator of Yatracos (1985), we do not require Θ to be discretised into a grid.

In spite of this, we can show strong consistency of θ_n with respect to $\theta^* = \operatorname{argmin}_{\theta \in \Theta} \operatorname{TVD}(p, p_\theta)$ under mild differentiability conditions (See Assumption 1, Appendix 3.A).

Proposition 4. *Suppose Assumption 1 holds. If $\theta_n = \operatorname{argmin}_{\theta \in \Theta} \operatorname{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$ is almost surely unique for all n large enough and the conditions of Proposition 2 hold, $\theta_n \xrightarrow{a.s.} \theta^*$ as $n \rightarrow \infty$. Similarly, if $\theta_n = \operatorname{argmin}_{\theta \in \Theta} \operatorname{TVD}(\hat{p}_{n,h_n}, \hat{p}_{\theta,n})$ is almost surely unique for all n large enough and the conditions of Proposition 3 hold, $\theta_n \xrightarrow{a.s.} \theta^*$ as $\min_{y \in \mathcal{Y}} n_y \rightarrow \infty$.*

Proof sketch. We give the proof for the discrete case only, as the continuous case follows similar arguments. First, we show that

$$0 \leq \operatorname{TVD}(p, p_{\theta_n}) - \operatorname{TVD}(p, p_{\theta^*}) \leq 2 \cdot \operatorname{TVD}(p_n, p).$$

Further, $\operatorname{TVD}(\hat{p}_n, p)$ goes to zero almost surely as $n \rightarrow \infty$. Since we care about the limiting value of θ_n , we may confine the analysis to large enough n (say $n > N$) for which $2 \cdot \operatorname{TVD}(\hat{p}_n, p) < \varepsilon$. Because $\operatorname{TVD}(p, p_{\theta^*}) - \operatorname{TVD}(p, p_{\theta_n}) < \varepsilon$ implies that $\theta_n \in B_\varepsilon$, restricting attention to $n > N$ is equivalent to restricting attention to the compact set B_ε . Thanks to this and the finite gradients, $\operatorname{TVD}(p, p_{\theta^*}) - \operatorname{TVD}(p, p_{\theta_n})$ converges to zero uniformly over B_ε . Together with the fact that $\theta_n \in B_\varepsilon$ for $n > N$, this implies the result. \square

3.5 Nonparametric Learning

In the context of misspecification, choosing a prior over the parameter space Θ is conceptually unappealing almost by definition: On the one hand, we readily admit that the probability model f_θ is misspecified, meaning that the interpretation of ‘good’ values for θ is not straightforward. On the other hand, we are forced

into having our quantification of uncertainty depend on a prior belief over the parameter θ . As a consequence, the inferred uncertainties are not straightforwardly interpretable in a Bayesian sense.

To avoid these complications, we opt for a different strategy: Instead of choosing a prior over θ , we impose an uninformative prior *directly* on the data-generating mechanism. This approach has recently been advocated in a series of papers (Lyddon, Walker, and Holmes, 2018; Fong, Lyddon, and Holmes, 2019) under the name ‘Bayesian Nonparametric Learning’ (NPL) and has three main benefits: Firstly, it perfectly suits the misspecified setting as discussed above. Secondly, it is suitable for generalised Bayesian inference with arbitrary loss functions. Thirdly, it produces exact inferences at low computational cost. For completeness, we give a brief description of the core components and the corresponding inference algorithm.

Suppose that we have access to the true data-generating mechanism p rather than some sample $\{x_i, y_i\}_{i=1}^n$. In this case, we have *no need* for uncertainty. In fact, we could simply compute

$$\theta^* = \theta(p) = \operatorname{argmin}_{\theta \in \Theta} \operatorname{TVD}(p, p_\theta).$$

In practice of course, this is impossible. In fact, it is this very impossibility that necessitates both the Frequentist and Bayesian notions of uncertainty. Taking this observation to heart, Lyddon, Walker, and Holmes (2018) and Fong, Lyddon, and Holmes (2019) have asked: Instead of quantifying uncertainty about ‘good’ parameter values by placing a prior on θ , why not place a prior on p ? In its characterization of Bayesian uncertainty, this approach closely tracks the classical bootstrap (Efron, 1979) and its Bayesian counterpart (Rubin, 1981).

While placing a prior over p is easier said than done, Lyddon, Walker, and Holmes (2018) and Fong, Lyddon, and Holmes (2019) show that Dirichlet Processes (DPs) are a computationally efficient way of doing this. For a measure p_π over $\mathcal{X} \times \mathcal{Y}$ encoding our prior beliefs about p and a scalar α determining the strength of this belief, we follow the approach in Fong, Lyddon, and Holmes (2019) and define the following (nonparametric) Bayesian prior about the data-generating mechanism:

$$p \sim \operatorname{DP}(\alpha, p_\pi).$$

Because we have assumed that our observations have been generated independently and identically distributed, this choice of prior yields a conjugate closed-form DP posterior as

$$p|\{x_i, y_i\}_{i=1}^n \sim \text{DP}(\alpha + n, p_{\pi, n});$$

$$p_{\pi, n} = \frac{\alpha}{\alpha + n} \cdot p_{\pi} + \frac{1}{\alpha + n} \cdot \sum_{i=1}^n \delta_{(x_i, y_i)}. \quad (3.5)$$

This suggests sampling $\{p^{(i)}\}_{i=1}^B$ from $\text{DP}(\alpha + n, p_{\pi, n})$ and then computing a sample $\{\theta_i\}_{i=1}^B$, where $\theta_i = \theta(p^{(i)})$. To make sampling from a DP computationally feasible, some form of truncation limit T is required. Again, we follow the suggestions of Fong, Lyddon, and Holmes (2019), yielding the *Posterior Bootstrap Sampling* algorithm.

Algorithm 2 Posterior Bootstrap Sampling

Input: $\{x_i, y_i\}_{i=1}^n, \alpha, p_{\pi}, T$
for $j = 1, 2, \dots, B$ **do**
 draw pseudo-samples $(\tilde{y}_i^{(j)}, \tilde{x}_i^{(j)})_{1:T} \stackrel{i.i.d.}{\sim} p_{\pi}$
 draw $(w_{1:n}^{(j)}, \tilde{w}_{1:T}^{(j)}) \sim \text{Dir}(1, \dots, 1, \alpha/T, \dots, \alpha/T)$
 get $p^{(j)} = \sum_{i=1}^n w_i^{(j)} \delta_{(y_i, x_i)} + \sum_{k=1}^T \delta_{(\tilde{y}_k^{(j)}, \tilde{x}_k^{(j)})}$
 get $p_{\theta, n}^{(j)}(x, y) = p_{\theta}(y|x) \left[\sum_{y \in \mathcal{Y}} p^{(j)}(x, y) \right]$
 Compute $\theta^{(j)} = \text{argmin}_{\theta \in \Theta} \text{TVD}(p^{(j)}, p_{\theta, n}^{(j)})$
return posterior bootstrap sample $\theta^{(1:B)}$

This Bayesian inference scheme has three rare and desirable properties: It is simple, embarrassingly parallel, and produces independent parameter samples.

There are two main levers for tuning the above algorithm: The prior p_{π} and the scalar α . Throughout our experiments, we use the limiting case of $\alpha \rightarrow 0$, which automatically eliminates the need to specify p_{π} (see Equation (3.5)). As pointed out by Fong, Lyddon, and Holmes (2019), this has the interpretation of positing a maximally uninformative prior belief about p . Computationally, the algorithm is equivalent to a generalised form of the Bayesian Bootstrap (Rubin, 1981) introduced by Lyddon, Holmes, and Walker (2019).

This choice of α is thus justified from a conceptual as well as a practical standpoint. Conceptually, it reflects the fact that we have no clear idea about the nature of p —since if we had, we could specify a model family $\{f_{\theta} : \theta \in \Theta\}$ that is not drastically misspecified. On a practical level, it simplifies the

inference algorithm by eliminating two hyperparameters (α and p_π) and leads to considerable speedups.

A more implicit lever for tuning the algorithm is the sub-routine one chooses to find the minimiser $\theta^{(j)}$. Minimising $\text{TVD}(p^{(j)}, p_{\theta,n}^{(j)})$ is generally very difficult: The function will not be convex in θ everywhere. Worse still, it will in fact be equal to its upper bound for most values of θ . This makes it crucial to find good initial values from which to start a gradient-based optimization process. To address this, we compute the minimisers by using maximum likelihood estimates as initialisers and then compute the (possibly local) minima of $\text{TVD}(p^{(j)}, p_{\theta,n}^{(j)})$ using the Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm.

3.6 Experiments

We verify the performance of our method on a number of synthetic and real world data examples and well as our problem of interest, crime. First, we study two canonical synthetic data examples: An ε -contaminated Poisson model and a zero-inflated binomial model. We find that our method recovers parameter estimates that are close to those of the uncontaminated data generating process and improve out-of-sample predictive performance. Next, we investigate performance on real world data with two standard classification models: Probit regressions and Neural Networks. For both models and across all data sets, we find that our method improves out-of-sample predictive performance. Lastly, we test our method on a data set of reported sexual crimes and find that our inference strategy is more robust to random clustering of reports on specific dates than standard inference.

We consider three evaluation criteria which we explain in more detail in Section 3.6.1. In addition to performing inference with the TVD and the KLD we also employed a fully Bayesian approach using Stan (Stan Development Team, 2018) with 4 chains and 1,000 posterior draws resulting in 4,000 draws from the posterior distributions in total. With respect to all three evaluation criteria, the fully Bayesian approach is consistently out-performed by the TVD and the KLD. This already demonstrates that using the Bootstrap approach proposed by Lyddon, Holmes, and Walker (2019) and Fong, Lyddon, and Holmes (2019) entails a certain robustness because it incorporates uncertainty contained within the data. We only show the full results in Appendix 3.B for brevity.

For all experiments, we use python’s `statsmodels` package to fit the (weighted) maximum likelihood estimate. All code is publicly available at <https://github.com>.

`com/laravomfell/tvd_loss`. Whenever the optimization of the BFGS did not converge, we excluded the resulting samples of the algorithm. This happened only for the ε -contaminated Poisson model (due to the numerically instable double-exponential parametrisation of λ) and affected a negligibly small number of samples.

3.6.1 Evaluation criteria

In our simulated experiments we consider three evaluation criteria: 1. the absolute difference between our inferred parameters and the truth, 2. absolute error on a test data set and 3. the predictive likelihood.

For our synthetic data examples, we generate 100 data sets all according to the same data-generating mechanism. For the real-world data examples, we generate 50 random splits of the data. We then pool performance across resamples/splits as follows: Taking $\theta_m^{(j)}$ to be the j -th sample on the m -th artificially generated training data set or split, our plots of criterion 1 then show quantiles of

$$d_{i,m} = \frac{1}{B} \sum_{j=1}^B |\theta_m^{(j)} - \theta|. \quad (3.6)$$

Next, we show quantiles of absolute errors

$$e_{i,m} = \frac{1}{B} \sum_{j=1}^B |y_{i,m}^{(j)} - y_{i,m}|, \quad (3.7)$$

where $y_{i,m}$ is the outcome of the i -th observation in the m -th artificially generated data set and $y_{i,m}^{(j)}$ is the expected value of the distribution $p_{\theta_m^{(j)}}(\cdot|x_{i,m})$.

Lastly, we compute the predictive likelihoods for all experiments and show quantiles of

$$l_{m,i} = \frac{1}{B} \sum_{j=1}^B p_{\theta_m^{(j)}}(y_{i,m}|x_{i,m}), \quad (3.8)$$

where $(y_{i,m}, x_{i,m})$ is the i -th observation in the m -th artificially generated training data set or split.

3.6.2 ε -contamination

We consider an ε -contaminated Poisson model based on Equation (3.3): A proportion $(1 - \varepsilon)$ of the data come from a standard Poisson model while $\varepsilon \in (0, 1)$ of the data come from a contamination distribution. Throughout the experiments, we fix the mean of the Poisson distribution as $\lambda = 3$, the proportion $\varepsilon = 0.15$ and consider contamination in form of an offset by a constant k . In other words, we generate our data as

$$y_i \stackrel{i.i.d.}{\sim} \text{Poisson}(\lambda) + \text{Bernoulli}(\varepsilon) \cdot k \quad (3.9)$$

For a fixed value of k , we generate $m = 100$ data sets of 500 observations. Each of these is split into train and test data sets according to an 80:20 split. We then use Algorithm 2 to infer two misspecified Poisson models—misspecified because they do not take the contamination into account—by drawing $B = 1000$ samples, one based on minimising the KLD and one based on minimising the TVD.

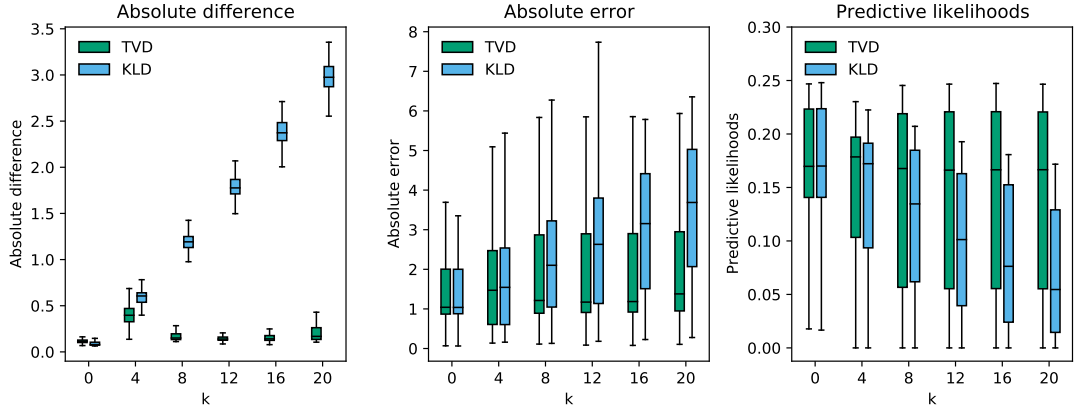


Figure 3.2: Difference in inference outcomes for the ε -contamination model of Equation (3.9) between using the **TVD** and the **KLD** as k is varied and $\varepsilon = 0.15$. **Left:** Absolute difference between inferred and true value of λ ; **Middle:** Absolute out-of-sample prediction error; **Right:** Predictive likelihood on out-of-sample data.

Figure 3.2 reports the results for our three key quantities: 1. the absolute difference between the true and inferred value of λ , 2. the absolute error on the test data and 3. the predictive likelihood.

As shown in the left panel of Figure 3.2, minimising the TVD produces parameters similar to those of the KLD in the absence of contamination ($k = 0$) and produces consistently better estimates as the degree of contamination becomes more severe. Similarly, the TVD improves out-of-sample prediction. Notably, its

median absolute error remains essentially constant. Lastly, Figure 3.2 also shows that the TVD consistently improves predictive likelihoods relative to the KLD—even under increasingly extreme contamination. This improvement in calibration is perhaps not surprising, as it is predicted by the theory outlined in Section 3.3.

3.6.3 Zero-inflation

We also consider a zero-inflated binomial regression model where the probability of success π out of $w = 8$ trials is modelled by a categorical covariate. Casting this in the language of Equation (3.3), this means that the contaminating distribution is a Dirac delta at zero. This implies the following model:

$$\begin{aligned} y_i | x_i &\stackrel{i.i.d.}{\sim} \text{Binomial}(8, \pi_i) \cdot (1 - \text{Bernoulli}(\varepsilon)) \\ \pi_i &= \text{logit}^{-1}(0.8 + 0.25x_i) \\ x_i &\stackrel{i.i.d.}{\sim} \text{Categorical}(\mathbf{p}, 4), \quad p_1 = \dots = p_4 = 1/4. \end{aligned}$$

Again, we generate $m = 100$ data sets of $n = 1000$ observations which we split into train and test data with an 80:20 split. We are interested in the same three quantities as in the ε -contamination example and report these in Figure 3.3. Instead of λ , we now report the absolute difference between the inferred and true value of $\beta = 0.25$. As before, the TVD guarantees parameter estimates closer to the uncontaminated data-generating process than those of the KLD. In turn, this yields superior out-of-sample prediction and predictive likelihoods.

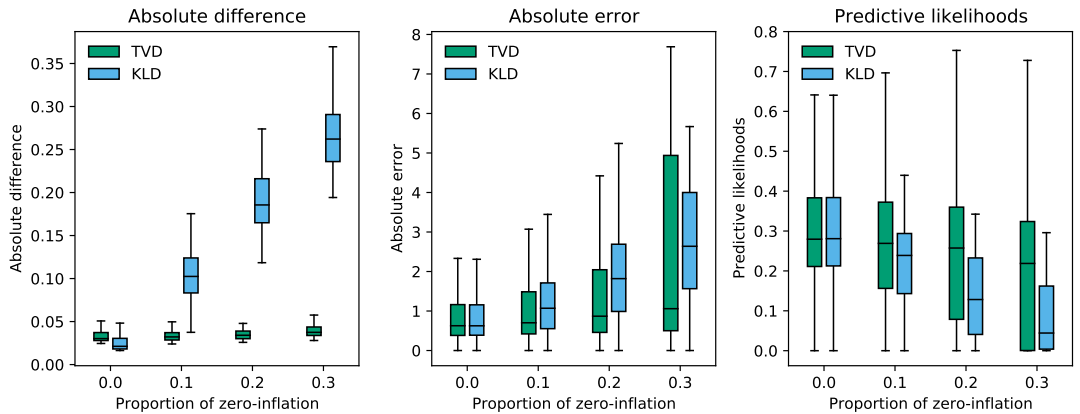


Figure 3.3: Difference in inference outcomes for the zero-inflation model between using the **TVD** and the **KLD** as the proportion ε of zeros is varied. **Left:** Absolute difference between inferred and true value of β ; **Middle:** Absolute out-of-sample prediction error; **Right:** Predictive likelihood on out-of-sample data.

3.6.4 Probit

The Probit model is a canonical statistical model for binary classification. For our evaluation, we select 5 data sets from the UCI repository with discrete co-variates: `mammograph`, `fourclass`, `heart`, `haberman` and `breast cancer`. For the `fourclass` data set, we test the first against the remaining classes.

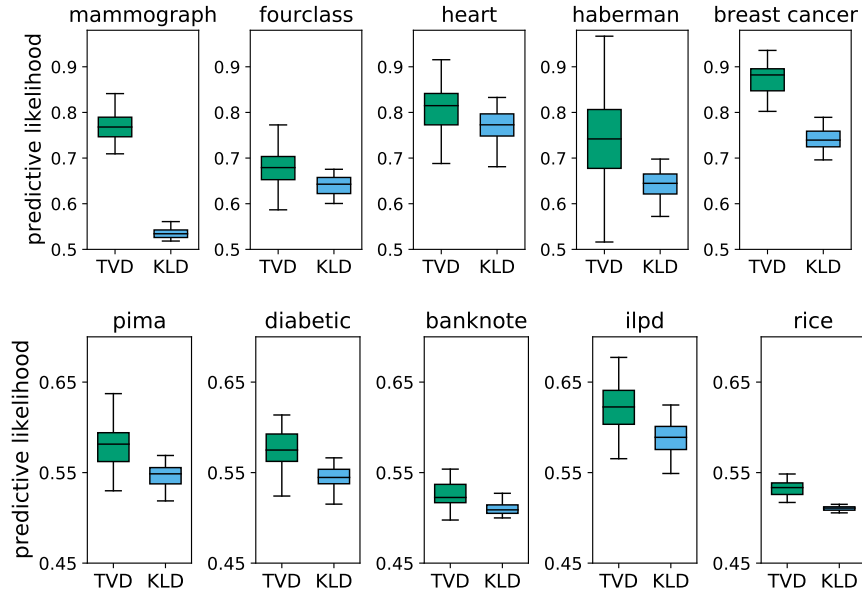


Figure 3.4: Predictive likelihoods for the Probit models (top) and single-layer Neural Networks (bottom).

For each data set, we produce $m = 50$ random splits of the data into training and test according to a 90:10 split. None of these data sets are easy to describe with simple off-the-shelf models which strongly suggests the presence of model misspecification. Our findings confirm this: Using the robust TVD produces a model with better predictive calibration than the KLD on all five data sets (see top row of Figure 3.4). While using the TVD yields a major improvement for the predictive likelihoods, the accuracy remains relatively similar to the KLD case, see also the top row of Figure 3.7.

3.6.5 Neural Network

Neural networks are another important Machine Learning classification model. Again, we select five data sets from UCI with binary outcomes: `pima`, `diabetic`, `banknote`, `ilpd` and `rice`. Unlike the data sets in the Probit section, these data sets have continuous covariates.

Again, we produce $m = 50$ random training and test splits with 90:10 split. We use stochastic gradient descent within `Pytorch` (Paszke et al., 2019) to get an initialiser for the Neural Network. Once again, the BFGS algorithm is used to find the TVD-minimising value of θ . All Neural Network examples were run with a single hidden layer of 50 nodes.

In spite of the added complication of continuous covariates, the bottom row of Figure 3.4 shows that we find similar improvements in the predictive calibration as for the Probit models.

3.6.6 Modelling crime incidence

Finally, we showcase our method on crime data. Figure 3.5a shows the police-recorded daily incidence of rapes between 2010 and 2017 in a major English police force area. The number on incidents is extremely high on the first of January of each year. We presume that this is due to the reporting of historic cases. If crime victims cannot remember the precise day on which a crime occurred, the crime record will state the first of the month or the first day of the year as the date of the crime instead. Sexual offences are a particularly stark example of this phenomenon since often a long time has passed between the first incidence of the crime and the victim reporting. As the time series of reports gets closer to the present, this ‘spikiness’ of reports decreases.

If we are now interested in inferring the daily incidence of rapes, then these high-report days can distort our inference. We demonstrate this issue in Figure 3.5b: We have to truncate the histogram at 22, since the high-report days lie so far away from the majority of the data that they do not fit on the same scale. At the same time, inference using the KLD is affected by these outlying days while our robust approach using the TVD is robust to those outliers. Clearly, now that we can already presume a generating process for our data, we could use a more specialized/bespoke model for this data. However, crime data is generally noisy and we often do not know the source of this noisiness. This example is meant to show how a generic robust inference strategy can improve inference in settings where crime counts are low, variable and the result of unknown recording choices.

3.7 Conclusion

We propose a new generalised Bayesian inference method using the Total Variation Distance (TVD) to robustify the inference of models for discrete-valued

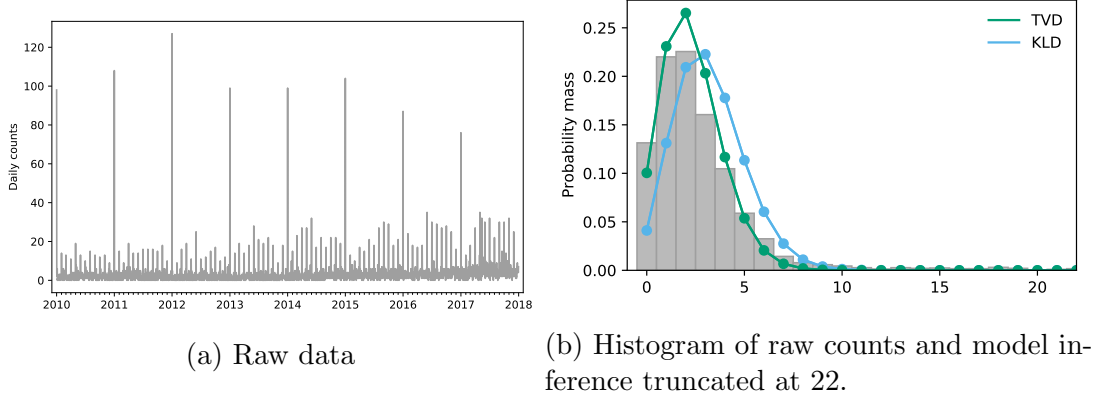


Figure 3.5: Incidence of sexual offences and model inference.

outcomes. The resulting inference procedure is based on an estimator of the TVD which possesses a range of desirable theoretical properties. In practice, our method yields significantly improved inference outcomes under model misspecification. This is especially valuable for complex applied problems such as the modelling of noisy crime data. Since we found no significant loss of efficiency in the absence of misspecification, we conclude that our method is a useful generic robustness approach.

Appendix 3.A Proofs

We give a detailed account for the concentration inequalities as well as related results derived in the main paper.

Proposition 1

Proof. One simply observes that by definition,

$$\begin{aligned}
 |\text{TVD}(c, p_\theta) - \text{TVD}(p_{\underline{\theta}}, p_\theta)| &= |(1 - \varepsilon)\text{TVD}(p_{\underline{\theta}}, p_\theta) + \varepsilon\text{TVD}(q, p_\theta) - \text{TVD}(p_{\underline{\theta}}, p_\theta)| \\
 &\leq |\varepsilon\text{TVD}(q, p_\theta) + \varepsilon\text{TVD}(p_{\underline{\theta}}, p_\theta)| \leq 2\varepsilon,
 \end{aligned}$$

which completes the proof. □

Corollary 1

Proof. We only give proof for the discrete case, as the arguments are the same for the continuous case. Denoting $A_n = \{|\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| > \varepsilon\}$,

$$\sum_{i=1}^{\infty} \mathbb{P}(A_n) \leq (2^{K_y + K_x + 1} - 2^2) \sum_{i=1}^n e^{-n\varepsilon^2/2} < \infty.$$

This immediately implies that we can apply the Borel-Cantelli Lemma to conclude that

$$\mathbb{P}(\lim_{n \rightarrow \infty} A_n) \leq \mathbb{P}(\limsup_{n \rightarrow \infty} A_n) = 0.$$

As ε was chosen arbitrarily, the proof is complete. \square

Corollary 2

Proof. We only give proof for the discrete case, as the arguments are the same for the continuous case. By Corollary 1, we know that there exists N such that with probability one, $|\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n}) - \text{TVD}(c, p_\theta)| \leq \eta$ for all $n \geq N$. Thus, for $n \geq N$ it holds that

$$\begin{aligned} & |\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n}) - \text{TVD}(p_\theta, p_\theta)| \\ & \leq |\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n}) - \text{TVD}(c, p_\theta)| + |\text{TVD}(p_\theta, p_\theta) - \text{TVD}(c, p_\theta)| \\ & \leq \eta + 2\varepsilon, \end{aligned}$$

which completes the proof. \square

Proposition 2

Proof. The proof proceeds in two steps. First, we bound $\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n})$ from above and below using the same terms. Second, we show that only one of these terms (namely $\mathbb{E}_{p^x}[\text{TVD}(p^{y|x}, f_\theta)]$) does not satisfy an exponential concentration towards zero.

For simplicity, we first give the proof for countable spaces \mathcal{X} and then extend it to uncountable Euclidean spaces \mathcal{X} later. Writing $p_\theta(x, y) = p_\theta(x|y)p^x(x)$ as

well as $\widehat{p}_{\theta,n}(x, y) = f_{\theta}(x|y)\widehat{p}_n^x(x)$, we begin by noting that

$$\text{TVD}(\widehat{p}_n, \widehat{p}_{\theta,n}) = \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |\widehat{p}_n(x, y) - \widehat{p}_{\theta,n}(x, y)|.$$

Defining further the functions $\Delta(x, y) = \widehat{p}_n(x, y) - p(x, y)$, $\Delta_{\theta,*}(x, y) = p(x, y) - p_{\theta}(x, y)$ and $\Delta_{\theta,n}(x, y) = p_{\theta}(x, y) - \widehat{p}_{\theta,n}(x, y)$, we can rewrite and lower bound the above as

$$\begin{aligned} & \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |\Delta(x, y) + \Delta_{\theta,*}(x, y) + \Delta_{\theta,n}(x, y)| \\ & \geq \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \left| |\Delta_{\theta,*}(x, y)| - |\Delta(x, y) + \Delta_{\theta,n}(x, y)| \right| \\ & \geq \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \left| |\Delta_{\theta,*}(x, y)| - |\Delta(x, y)| - |\Delta_{\theta,n}(x, y)| \right| \\ & \geq \left| \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |\Delta_{\theta,*}(x, y)| - |\Delta(x, y)| - |\Delta_{\theta,n}(x, y)| \right| \\ & = |\text{TVD}(p, p_{\theta}) - \text{TVD}(\widehat{p}_n, p) - \text{TVD}(p, \widehat{p}_{\theta,n})|. \\ & \geq \text{TVD}(p, \widehat{p}_{\theta}) - [\text{TVD}(\widehat{p}_n, p) + \text{TVD}(p_{\theta}, \widehat{p}_{\theta,n})]. \end{aligned}$$

The upper bound uses the decomposition and is a direct consequence of the triangle inequality. Specifically,

$$\begin{aligned} & \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |\Delta(x, y) + \Delta_{\theta,*}(x, y) + \Delta_{\theta,n}(x, y)| \\ & \leq \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (|\Delta(x, y)| + |\Delta_{\theta,*}(x, y)| + |\Delta_{\theta,n}(x, y)|) \\ & = \text{TVD}(p_*, p_{\theta,*}) + [\text{TVD}(p_n, p_*) + \text{TVD}(p_{\theta,*}, p_{\theta,n})]. \end{aligned}$$

Noting the form of the upper and lower bounds together with the fact that

$$\text{TVD}(p, p_{\theta}) = \mathbb{E}_{p^x} [\text{TVD}(p^{y|x}, f_{\theta})],$$

it becomes clear that finding concentration inequalities for both $\text{TVD}(\widehat{p}_n, p)$ and $\text{TVD}(p_{\theta}, \widehat{p}_{\theta,n})$ towards zero suffices to prove the desired result. By assumption, it also holds that the joint distributions p and \widehat{p}_n have an alphabet of size at most $K_x \cdot K_y$. Numerous exponential concentration inequalities apply to this setting. While stronger results are available for the case where n is small relative to $K_x + K_y$ (Mardia et al., 2019), we rely on Theorem 2.1 of (Weissman et al.,

2003) for simplicity and since the rates in n remain the same. The latter shows that

$$\begin{aligned}\mathbb{P}(\text{TVD}(p, \hat{p}_n) > \varepsilon) &\geq \delta(n, \varepsilon, K_x + K_y) \\ \delta(n, \varepsilon, K) &= (2^K - 2) \cdot e^{-n\varepsilon^2/2}\end{aligned}$$

Further, we also have that for any $\theta \in \Theta$,

$$\begin{aligned}\text{TVD}(p_\theta, \hat{p}_{\theta,n}) &= \frac{1}{2} \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} |p^x(x) - \hat{p}_n^x(x)| f_\theta(y|x) \\ &= \frac{1}{2} \sum_{x \in \mathcal{X}} |p^x(x) - \hat{p}_n^x(x)| \underbrace{\sum_{y \in \mathcal{Y}} f_\theta(y|x)}_{\leq 1} \\ &\leq \text{TVD}(\hat{p}_n^x, p^x).\end{aligned}$$

Since \hat{p}_n^x and p^x have an alphabet of size at most K_x , the same type of concentration inequality applies here, too so that

$$\mathbb{P}(\text{TVD}(p^x, \hat{p}_n^x) > \varepsilon) \leq \delta(n, \varepsilon, K_x)$$

Setting now $\varepsilon = 2/\eta$ and using a union bound argument, we find that

$$\begin{aligned}\mathbb{P}(\text{TVD}(p, \hat{p}_n) + \text{TVD}(p^x, \hat{p}_n^x) > \varepsilon) &\leq \delta(n, 2/\eta, K_x) + \delta(n, 2/\eta, K_x + K_y) \\ &\leq 2\delta(n, 2/\eta, K_x + K_y)\end{aligned}$$

Since by virtue of our previous derivations we also have that

$$|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| \leq \text{TVD}(p, p_n) + \text{TVD}(p^x, \hat{p}_n^x),$$

this completes the proof. □

Proposition 3

Proof. The first part of the proof follows exactly like in the discrete case. The only difference becomes the replacement of the summation $\sum_{x \in \mathcal{X}}$ with an integration operation. Defining the joint, marginal and conditional kernel density estimates

as

$$\begin{aligned}
\widehat{p}_{n,h_n}(x,y) &= \frac{1}{n} \sum_{i=1}^n \delta_{y_i}(y) \cdot \frac{1}{h_n^d} K\left(\frac{x_i - x}{h_n}\right) \\
\widehat{p}_{n,h_n}^x(x) &= \sum_{y \in \mathcal{Y}} \widehat{p}_{n,h_n}(x,y) \\
\widehat{p}_{n,h_n}^y(y) &= \int_{\mathcal{X}} \widehat{p}_{n,h_n}(x,y) dx \\
\widehat{p}_{n,h_n}^{x|y}(x|y) &= \widehat{p}_{n,h_n}(x,y) / \widehat{p}_{n,h_n}^y(y) \\
p_{n,h_n}^{y|x}(y|x) &= \widehat{p}_{n,h_n}(x,y) / \widehat{p}_{n,h_n}^x(x).
\end{aligned}$$

Further, we define $\widehat{p}_{\theta,n,h_n}(x,y) = f_{\theta}(x|y)\widehat{p}_{n,h_n}^x(x)$ as the same hybrid-type distributions that were used in the proof for the discrete case. Using now the same basic inequalities as before, we find that

$$|\text{TVD}(p_{n,h_n}, \widehat{p}_{\theta,n,h_n}) - \text{TVD}(p, p_{\theta})| \leq \text{TVD}(p_{n,h_n}, p) + \text{TVD}(p_{\theta}, \widehat{p}_{\theta,n,h_n}).$$

Similarly, we can use the same arguments as in the discrete case to conclude that for any $\theta \in \Theta$,

$$\text{TVD}(p_{\theta}, \widehat{p}_{\theta,n,h_n}) \leq \text{TVD}(\widehat{p}_{n,h_n}^x, p^x).$$

Notice that by definition, \widehat{p}_{n,h_n}^x is just a regular kernel density estimate for an absolutely continuous density. Thus, we can apply Theorem 1 (Chapter 3) in Devroye and Györfi (1985) to conclude that $\text{TVD}(p_{\theta}, \widehat{p}_{\theta,n,h_n})$ goes to zero exponentially fast.

Next, we use the triangle inequality to conclude that for $q(x,y) = \widehat{p}_{n,h_n}^{x|y}(x|y)p^y(y)$,

$$\text{TVD}(\widehat{p}_{n,h_n}, p) \leq \text{TVD}(\widehat{p}_{n,h_n}, q) + \text{TVD}(q, p).$$

Since it holds that

$$\begin{aligned}
\text{TVD}(\widehat{p}_{n,h_n}, q) &= \frac{1}{2} \sum_{y \in \mathcal{Y}} \int_{\mathcal{X}} \widehat{p}_{n,h_n}^{x|y}(x|y) |\widehat{p}_{n,h_n}^y(y) - p^y(y)| dx \\
&= \frac{1}{2} \sum_{y \in \mathcal{Y}} |\widehat{p}_{n,h_n}^y(y) - p^y(y)| \underbrace{\int_{\mathcal{X}} \widehat{p}_{n,h_n}^{x|y}(x|y) dx}_{=1, \text{ for all } y} \\
&= \text{TVD}(\widehat{p}_{n,h_n}^y, p^y).
\end{aligned}$$

Notice that regardless of h_n , we actually have that $\widehat{p}_{n,h_n}^y = \widehat{p}_n^y = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}(y)$,

so that the same concentration inequalities in Weissman et al. (2003) already applied in the proof of the discrete case also hold for this last expression. For the second term resulting from applying the triangle inequality, we have that

$$\begin{aligned}\text{TVD}(q, p) &= \frac{1}{2} \sum_{y \in \mathcal{Y}} \underbrace{p^y(y)}_{\leq 1} \int_{\mathcal{X}} |\hat{p}_{n, h_n}^{x|y}(x|y) - p^{x|y}(x|y)| dx \\ &\leq \sup_{y \in \mathcal{Y}} \text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y)).\end{aligned}$$

Now note that by definition, for a fixed value of y , $\hat{p}_{n, h_n}^{x|y}$ is a regular kernel density estimate based on $n_y = \sum_{i=1}^n \delta_y(y_i)$ observations. Thus, $\text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y))$ satisfies a concentration inequality for any fixed y . We can use this knowledge as follows: Define the random variable $\tilde{y} : \mathbb{N} \rightarrow \mathcal{Y}$ to be drawn uniformly from the set of maximizers

$$\arg \max_{y \in \mathcal{Y}} \text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y))$$

and note that by Theorem 1 in Chapter 3 of Devroye and Györfi (1985)¹ we have a *conditional* upper-bound

$$\mathbb{P}\left(\max_{y \in \mathcal{Y}} \text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y)) > \eta \middle| \tilde{y}(n) = z\right) \leq \delta(z);$$

where for $n_z = \sum_{i=1}^n \delta_z(y_i)$ the number of times that $y_i = z$ occurred in the sample we have

$$\delta(z) = 3e^{-n_z \cdot \eta^2 / 50} + e^{-2n_z \eta^2 / 25}.$$

By taking $h : \mathcal{Y} \rightarrow [0, 1]$ to be the true (and unknown) probability mass function of \tilde{y} and summing over the conditional upper bound, we obtain that

$$\begin{aligned}\mathbb{P}\left(\max_{y \in \mathcal{Y}} \text{TVD}(\hat{p}_{n, h_n}^{x|y}(\cdot|y), p^{x|y}(\cdot|y)) > \eta\right) &\leq \sum_{z \in \mathcal{Y}} h(z) \delta(z) \\ &\leq \max_{z \in \mathcal{Y}} \delta(z) \\ &= \delta_1 = \min_{z \in \mathcal{Y}} \left\{ 3e^{-n_z \cdot \eta^2 / 50} + e^{-2n_z \eta^2 / 25} \right\}\end{aligned}$$

¹Though the constant is not stated explicitly in the original Theorem, one can work it out by tracing the relevant steps of Lemma 2 in the same Chapter and collecting the bounds. Doing so is tedious and yields a complicated expression which we have given an upper bound for here.

Based again on Theorem 1, Chapter 5 in Devroye and Györfi (1985), one also finds that for a kernel density estimate \hat{p}_{n,h_n}^x based on n data points,

$$\mathbb{P}(\text{TVD}(\hat{p}_{n,h_n}^x, p^x) > \eta) \leq \delta_2 = 3e^{-n\cdot\eta^2/50} + e^{-2n\eta^2/25}.$$

Reusing the discrete concentration inequalities from before, we also have that

$$\mathbb{P}(\text{TVD}(\hat{p}_{n,h_n}^y, p^y) > \eta) \leq \delta_3 = (2^{K_y} - 2) \cdot e^{-n\cdot\eta^2/2}.$$

To obtain the union bound (and thereby the desired result), we now set $\eta = \varepsilon/3$ to conclude that

$$\begin{aligned} \mathbb{P}(|\text{TVD}(\hat{p}_{n,h_n}, \hat{p}_{\theta,n,h_n}) - \text{TVD}(p, p_\theta)| > \varepsilon) &\leq \delta_1 + \delta_2 + \delta_3 \\ &= \min_{z \in \mathcal{Y}} \left\{ 3e^{-n_z \cdot \varepsilon^2/450} + e^{-2n_z \varepsilon^2/225} \right\} \\ &\quad + 3e^{-n \cdot \varepsilon^2/450} + e^{-2n \varepsilon^2/225} \\ &\quad + (2^{K_y} - 2) \cdot e^{-n \cdot \varepsilon^2/18}, \end{aligned}$$

which completes the proof. □

Proposition 4

Assumption 1. θ^* is unique and or some $\varepsilon > 0$,

$$B_\varepsilon = \{\theta : |\text{TVD}(p, p_\theta) - \text{TVD}(p, p_{\theta^*})| < \varepsilon\}$$

is compact. Further, $\text{TVD}(p, p_\theta)$ and $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$ are continuous on B_ε . Lastly, $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$ has finite gradients (with respect to θ) on B_ε for all n large enough in the sense that the finiteness condition

$$\limsup_{n \rightarrow \infty} \sup_{\theta' \in B_\varepsilon} \frac{\partial}{\partial \theta} \text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) \Big|_{\theta=\theta'} < \infty$$

holds almost surely.

Proof. We give the proof for the discrete case only, as the continuous case follows similar arguments. Roughly speaking, the proof proceeds in three steps: First, we show that for large enough n , $\theta_n \in B_\varepsilon$. As this implies that we can confine the analysis to a compact set, it drastically simplifies the subsequent analysis. Second, we prove that over B_ε , $|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)|$ converges almost

surely and *uniformly* to zero. Third, we use a standard argument to conclude that $\theta_n \xrightarrow{a.s.} \theta^*$.

First, observe that

$$\begin{aligned} \text{TVD}(p, p_{\theta_n}) &\leq \text{TVD}(\hat{p}_n, p_{\theta_n}) + \text{TVD}(\hat{p}_n, p) \\ &\leq \text{TVD}(\hat{p}_n, p_{\theta^*}) + \text{TVD}(\hat{p}_n, p) \\ &\leq \text{TVD}(p, p_{\theta^*}) + 2 \cdot \text{TVD}(\hat{p}_n, p), \end{aligned}$$

where the first and last lines follow by the triangle inequality, while the second line follows by definition of θ_n and θ^* . Further, note that by definition of θ^* ,

$$\text{TVD}(p, p_{\theta^*}) \leq \text{TVD}(p, p_{\theta_n}).$$

Combining these two inequalities,

$$0 \leq \text{TVD}(p, p_{\theta^*}) - \text{TVD}(p, p_{\theta_n}) \leq 2 \cdot \text{TVD}(\hat{p}_n, p).$$

Applying Theorem 2.1 of Weissman et al. (2003), we know that $\text{TVD}(\hat{p}_n, p)$ converges to zero in probability exponentially fast. By the Borell-Cantelli argument already used for the proof of Corollary 1, this implies that $\text{TVD}(\hat{p}_n, p) \xrightarrow{a.s.} 0$. Hence, for any $\xi > 0$ there is N so that for $n \geq N$ and almost surely,

$$0 \leq \text{TVD}(p, p_{\theta_n}) - \text{TVD}(p, p_{\theta^*}) \leq \xi.$$

Choosing $\xi = \varepsilon$, we can conclude that for $n \geq N$, $\theta_n \in B_\varepsilon$ (almost surely).

In the second step, we use the fact that we can restrict our analysis to $n \geq N$ (i.e., to B_ε) in order to prove uniform convergence. Recall that by Corollary 1, we have pointwise convergence: for each $\theta \in B_\varepsilon$, $|\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| \xrightarrow{a.s.} 0$. Further, $\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$ is strongly stochastically equicontinuous for $\theta \in B_\varepsilon$. This follows by a standard argument using the Mean Value Theorem in conjunction with the finiteness assumption on its gradients. Specifically, one can use the reasoning outlined on p. 340 in Davidson (1994), Equations (21.55) – (21.57) to conclude that the function is strongly stochastically equicontinuous by Theorem 21.10 of the same text. By Theorem 21.8 of the same text, this together with pointwise convergence implies that $\sup_{\theta \in B_\varepsilon} |\text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) - \text{TVD}(p, p_\theta)| \xrightarrow{a.s.} 0$.

The third and last step now consists in showing that as desired, $\theta_n \xrightarrow{a.s.} \theta^*$. This follows immediately by applying a standard result, see e.g., Lemma 2 in Ye and Johnson (2001). Notice that we can apply this result in spite of Θ being

non-compact: As we only care about the limit, we can re-cast the minimisation as occurring over a compact space. In particular,

$$\min_{\theta \in \Theta} \text{TVD}(\hat{p}_n, \hat{p}_{\theta,n}) = \min_{\theta \in B_\varepsilon} \text{TVD}(\hat{p}_n, \hat{p}_{\theta,n})$$

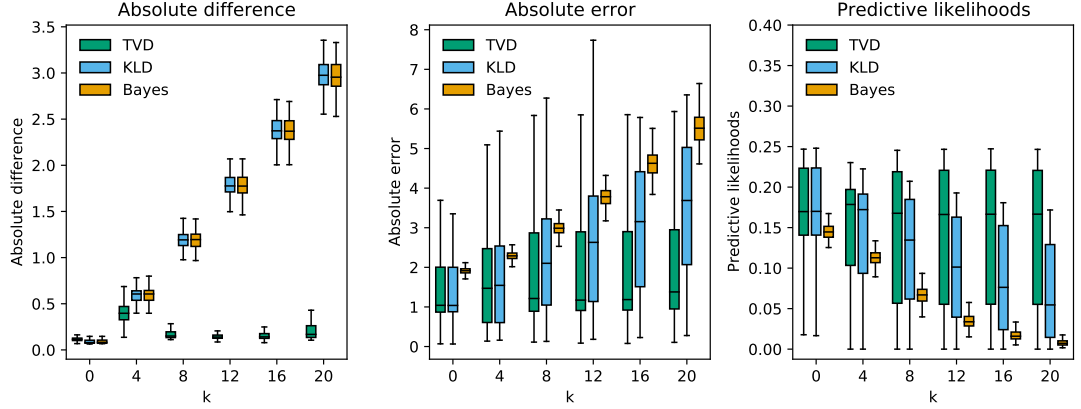
for all $n \geq N$.

The proof for the continuous case proceeds along the same lines once one replaces \hat{p}_n with \hat{p}_{n,h_n} . The only complication is the exponential concentration inequality: Instead of Theorem 2.1 of Weissman et al. (2003), one needs to use the results of Devroye and Györfi (1985) together with the arguments made for the proof of Proposition 3. \square

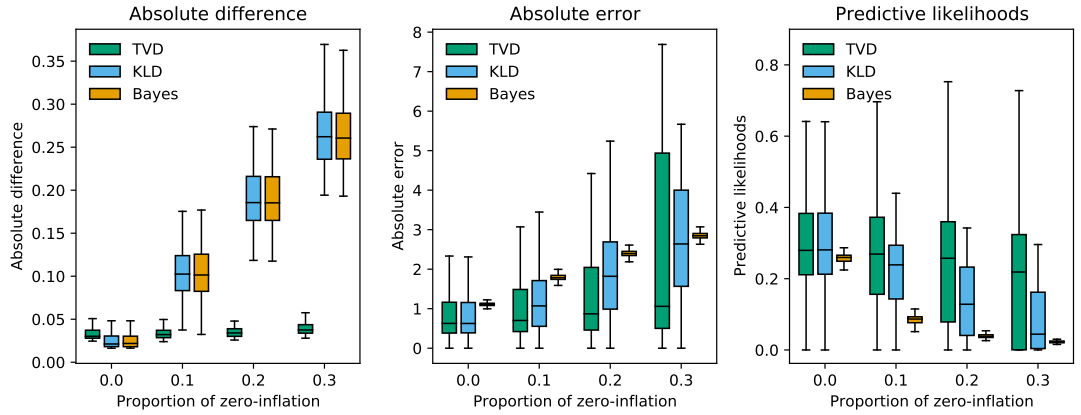
Appendix 3.B Full experimental results

The fully Bayesian approach using `stan` did not perform competitively relative to any method using NPL. For completeness, we report them here for the two synthetic data examples and all three evaluation criteria.

As for the Probit case, the accuracy remains relatively stable when the KLD is replaced by the TVD, see also the bottom row of Figure 3.7.



(a) Left: Absolute difference between estimated $\hat{\lambda}$ and true data-generating $\lambda = 3$ for different values of k with $\varepsilon = 0.15$ and $n_{\text{train}} = 400$. Middle: Absolute out-of-sample prediction error for different values of k with $n_{\text{test}} = 100$. Right: Predictive likelihood on out-of-sample test data ($n_{\text{test}} = 100$) for different values of k .



(b) Absolute difference between estimated $\hat{\beta}$ and true data-generating $\beta = 0.25$ under increasing zero-inflation with $n_{\text{train}} = 800$. Absolute out-of-sample prediction error ($n_{\text{test}} = 200$) under increasing contamination. Predictive likelihood on out-of-sample test data ($n_{\text{test}} = 200$) with increasing contamination

Figure 3.6: Full simulation results comparing inference using the KLD, the TVD and a fully Bayesian approach.

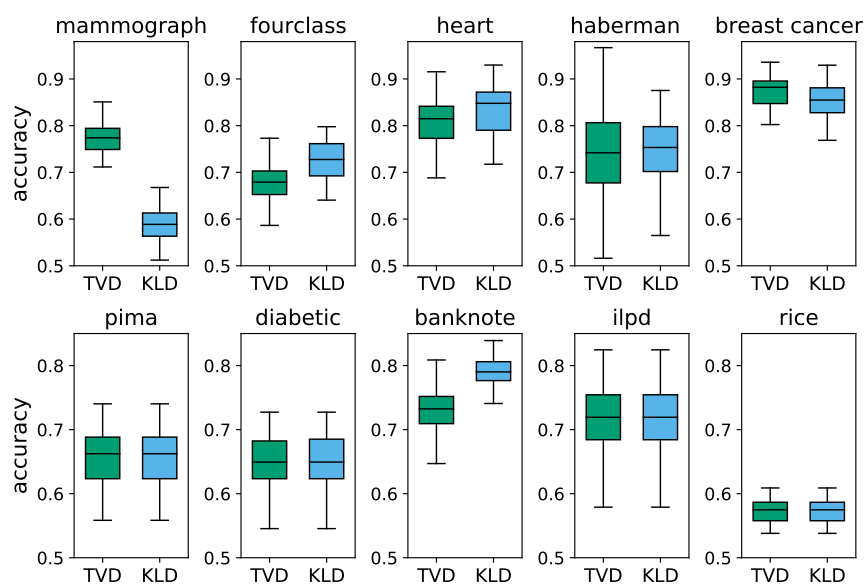


Figure 3.7: Predictive accuracy from 50 random splits for Probit models (top) and Neural Networks (bottom).

Conclusion

This thesis consists of three independent research studies on ethnic bias in policing, the reporting of domestic abuse and a new robust inference strategy. I will first summarise the chapters in a summary section before discussing the challenges of modelling observational behavioural data. Finally, I will discuss how analytical insights can be translated into policy. The work and discussion presented here seeks to cast a new light on the intersection between data and statistical modelling in complex applied behavioural science research.

Summary

In Chapter 1, we model police officers' search behaviour as the result of composite processes within the police and local neighbourhoods and of experiences with crime suspects. Following a panel of 36,000 searches by 1,100 police officers at the West Midlands police force, we provide officer-specific measures of officer bias defined as over-searching. Over-searching means that an ethnic groups makes up a larger share in the officer's searches than in a suitable baseline. In our study, we explore two baselines: the ethnic composition of crime suspects officers interact with and the ethnic composition of the areas they patrol.

The crime suspect baseline addresses frequent claims put forth by police that stop and search rates—which disproportionately target Black and Asian people—simply reflect crime figures. The patrolling baseline compares an officer's searches to the group most immediately affected by the officer's decisions: the people in the officer's patrolling beat. We show that the vast majority of officers over-search ethnic minorities against both baselines. More precisely, we find that virtually all officers over-search Asian people relative to their suspect baseline. For the patrolling baseline, we observe an interesting split where one half of officers over- and the other half under-searches Asian people. For Black searches, we find that more officers over-search Black people relative to their suspect baselines than

under-search. However, for the patrolling baseline, virtually all police officers in our sample over-search Black people. The vast majority of police officers in our sample under-search White people, irrespective of which baseline we look at.

However, limiting ourselves to an analysis of bias only at the police-officer level misses the fact that there are layers of bias in policing: We show that decisions at various levels of the policing structure contribute to the bias experienced by ethnic minorities. Over-patrolling of minority areas is a key factor. This finding is intuitive: Police officers are sent into more ethnically diverse areas. There, they then over-search ethnic minorities as they encounter them on patrol.

Decomposing the overall search bias at the police force level, we find that the over-representation of Asian people in stop and search is primarily accounted for by over-patrolling. In contrast, the over-representation of Black people is a combination of officer bias and over-patrolling effects, with the larger contribution coming from officers' biases.

There are aspects to studying ethnic bias in policing that our study does not cover. For example, we highlight the relevance of institutional layers of bias and their interaction. At the same time, the officer team layer is relatively coarsely operationalised. This is because the data available from West Midlands Police only provided information about team membership, not about the role within that team or supervisor relationships. Other work by Johnson, Tress, et al. (2019) and Quispe-Torreblanca and Stewart (2019) has shown that team composition, structure and responsibilities have tremendous effect on the dynamics within a team and the behaviour of its officers—a phenomenon partially shown in our results on team composition. Future work with access to similar data should explore this layer in more detail. Furthermore, the summer of 2020 also marks a fundamental shift in the prevalence of racism as a topic in public conversations. Following the deaths of Breonna Taylor, George Floyd and many other Black and Brown individuals at the hands of police, large-scale Black Lives Matter protests around the world led many public institutions, especially police forces, to grapple with their own racial and ethnic biases and histories.

Sudden public attention on racial bias in basketball refereeing following the publication of a study subsequently changed referees' decision-making (Pope, Price, and Wolfers, 2018). It is an open question if such effects would translate to policing. It should be noted that prior to the beginning of our study window (and our data coverage), West Midlands Police officers were part of various implicit bias trainings. Since we still document significant disparities in officer decision-making, we would like to point to multiple studies which note the limitations of

re-training as a solution to systemic biases (Onyeador, Hudson, and Lewis Jr, 2020; Roth and Sant’Anna, 2021; Sim, Correll, and Sadler, 2013; Smith, 2015).

In Chapter 2, we investigate whether spillover effects can occur in the reporting of crime. Typically, spillover effects are studied only in the context of offender behaviour, not on the victim’s part. We study this question in the particular context of domestic abuse, which exhibits multiple characteristics which make this an interesting environment: Domestic abuse is highly prevalent but seriously under-reported. Further, many victims of domestic abuse disclose their abuse to their social environment, even if not to police. Based on this, we study whether geographic proximity is conducive of spillover effects in decisions to report domestic abuse. Our data is the complete set of calls to police concerning domestic abuse in 2018 in a major English city. It stands to reason that the overlap between geographical proximity and social support networks (e.g., Osborne, Lau, and Britton, 2012) may result in localised spillover effects.

We study this question using a spatio-temporal Hawkes-type point process which can differentiate between clustering of reports in space and time on the one hand and spillovers on the other hand. Specifically, we check if the reporting of domestic abuse by one victim/survivor triggers further victims in the neighbourhood to report to police. We specifically seek to distinguish between spillovers due to initial reporting and the police response. We do so by extending an existing Hawkes process specification due to Zhuang and Mateu (2019) through the introduction of said distinction between spillover channels.

We find no evidence to support triggering in domestic abuse reporting. Any effects are extremely localized to a geographic range of 400m and 6 days after a report. Of the 6,084 reports of abuse in our data set, only 9.75×10^{-7} % are triggered by other events. Instead of triggering, we find highly periodic reporting patterns in line with other research on domestic abuse (Rotton and Cohn, 2001; Brimicombe and Cafe, 2012).

The work in Chapter 2 rests on the notion that spatial proximity accounts for some subset of social relations (see victim reporting to neighbors in Osborne, Lau, and Britton, 2012). Clearly, survivors’ social connections extend far beyond just their neighbourhoods. A recent paper by Fadlon and Nielsen (2019) finds that behaviour change following a health shock extends to immediate family and can potentially even alter coworkers’ behaviour. With more data about individuals’ social contacts and relations we could explore this question of potential spillovers and connect to a rich literature on spillover effects in networks (Galizzi

and Whitmarsh, 2019; Aronow et al., 2020).

The two applied studies presented in Chapters 1 and 2 are purely exploratory. A natural next step to exploring their generalisability is a pre-registered replication on new data (Van den Akker et al., 2019). We did not pre-register any studies presented here since many analysis decisions were taken conditional on the data observed. It is a well-known problem that such data-dependent decisions can affect the conclusions drawn from the research (Simmons, Nelson, and Simonsohn, 2011; Gelman and Loken, 2014). The studies are thus a useful first step in exploring the questions of ethnic bias and domestic abuse and will require further, pre-registered analyses.

In Chapter 3, we study the fundamental assumptions of likelihood-based statistical analysis such as maximum likelihood estimation or Bayesian inference. A crucial assumption underlying such methods is the idea of correct model specification. In other words, one assumes that the likelihood model used to represent the data-generating process can recover the underlying truth for a particular parametrization. While this assumption is often a useful approximation to how statistical analysis is conducted, it causes severe problems under even mild forms of misspecification. We propose a remedy for this problem by adapting the way information from a sample is related to the likelihood model in the context of discrete-valued outcomes.

Models of discrete-valued outcomes are easily misspecified if the data exhibit zero-inflation, overdispersion or contamination. Without additional knowledge about the existence and nature of this misspecification, model inference and prediction are adversely affected. Here, we introduce a robust discrepancy-based Bayesian approach using the Total Variation Distance (TVD). Our approach builds on prior work by Jewson, Smith, and Holmes (2018) which robustifies parameter inferences against the above mentioned forms of misspecification. It does so by replacing the information measure introduced by Kullback and Leibler (1951) with the TVD. In the process, we address and resolve two challenges: First, we study convergence and robustness properties of a computationally efficient estimator for the TVD between a parametric model and the data-generating mechanism. Note that an estimator is necessary because we precisely do not have access to the true data-generating process, but only our sample. Secondly, we provide an efficient inference method adapted from Lyddon, Holmes, and Walker (2019) which corresponds to formulating an uninformative nonparametric prior directly

over the data-generating mechanism. This is an appealing implementation because instead of quantifying uncertainty over the parameter, we instead quantify uncertainty over the object that we ultimately do not know, the data-generating process. Lastly, we empirically demonstrate that our approach is robust and significantly improves predictive performance on a range of simulated and real world data. In two simulation studies we find that, using our method, parameter inferences are unaffected by injections of substantial contamination. This also has a positive effect on the predictive capabilities of the models and we confirm this on a number of real-world data sets. Finally, we highlight the practical relevance of our method by inferring the incidence of sexual crimes.

The limitations of the work presented in Chapter 3 are less due to limited data but instead are inherent to our approach. Relative to the Kullback-Leibler Divergence (KLD), the TVD is a less efficient measure of information under correct specification and a biased estimator. Further, it is only an attractive replacement for the KLD in the case of discrete-valued outcomes. Conversely, other approaches such the Maximum Mean Discrepancy (Briol et al., 2019) cannot be used with discrete-valued outcomes. It would be more desirable to have a black-box approach capable of accommodating a wide range of outcomes and models. Finding more general but practically appealing robustness approaches which can be used in most circumstances remains a challenge that is yet to be resolved.

Discussion

Overall, the studies in this thesis emphasize that doing behavioural science based on field data is both a promising endeavour and difficult. Lab-based research allows for exploration of effects in tightly controlled environments. Researchers can use randomised manipulation of a wide range of factors to identify the key mechanisms behind a phenomenon (Mook, 1983; Fox, Erner, and Walters, 2015). At the same time, such research has often been criticized for trading off methodological rigour with external validity or generalisability (Black, 1955; Pruitt and Kimmel, 1977; Levitt and List, 2007a). The Merriam-Webster definition of behavio(u)r(al) science states that a goal of behavioural science is “to generalize about human behavior in society” (Merriam-Webster, 2021). A central question is how behaviours document in the lab factor into decision-making in ‘the wild’?

Research on decisions made in naturalistic environments and recorded as data can present a remedy to the generalisability question (Maner, 2016). At the same

time, a well-defined separation between field and lab studies is somewhat arbitrary as many studies skillfully blur the boundaries between the two by conducting lab-in-the-field studies (Harrison, Lau, and Williams, 2002; Charness and Villeval, 2009), by conducting experiments with the population of interest directly (Alevy, Haigh, and List, 2007), by studying naturalistic game settings (Van Dolder et al., 2015; Teeselink et al., 2020) or by exploring effects in both types of studies (Bohren, Imas, and Rosenberg, 2018; Galizzi and Navarro-Martinez, 2019; Folke and Rickne, 2020).

Indeed, many effects documented in the tightly controlled environment of a lab match behaviour in real-world conditions (e.g., Quispe-Torreblanca, Stewart, et al., 2019; Barberis, 2013; Mitchell, 2012; Alm, Bloomquist, and McKee, 2015; Galizzi and Navarro-Martinez, 2019). Often however, the translation is not straightforward and many effects do not materialise (Bryan et al., 2020). The reasons for that are many-fold and involve participant diversity (Henrich, Heine, and Norenzayan, 2010; Belot, Duch, and Miller, 2015), participants knowing they are part of a study, abstraction from real-world decision contexts and relatively unconstrained attention (Levitt and List, 2007b; Winking and Mizer, 2013).

But there is a further complication with the initial claim that lab-based research is a trade-off between generalisability and methodological rigour. The replication crisis has vividly shown that even the supposedly straightforward analysis of experimental data holds many traps. In addition to concerns about publication bias, outcome reporting bias and questionable research practices (Dwan et al., 2008; Fanelli, 2010; John, Loewenstein, and Prelec, 2012; Franco, Malhotra, and Simonovits, 2014), many more technical issues have led to concerns about the validity of science (Pashler and Wagenmakers, 2012). Some of these issues include low statistical power (Tressoldi, 2012), researcher degrees of freedom (Simmons, Nelson, and Simonsohn, 2011; Wagenmakers et al., 2012) and the dominance of the null hypothesis testing framework (Gigerenzer, 2004; Lakens, Scheel, and Isager, 2018; Scheel et al., 2020).

These issues are even more important in analyses of real-world data. In contrast to lab-based experiments, where the data take the form that the researcher intends them to take prior to data collection, good data on real decisions in the real world are not easy to come by. That is because often we do not have access to the data we would like or because the data do not exist. Unlike in the lab, researchers have no control over the data collection process. Instead, data is often generated by third parties. As a result, the data often do not contain all control variables of interest or need, decision contexts are missing and samples are

often non-random. Then, many difficulties arise such as Simpson’s paradox, the ecological fallacy (Freedman, 1999), heterogeneity (Heckman, 1991) and omitted variable bias. A full discussion of the various concerns around appropriate statistical modelling would unfortunately go beyond the scope of this thesis discussion. Here, I limit myself to just two aspects that were relevant for the studies presented here: 1. data-generating processes and 2. model specification.

1. Data-generating process Researchers have to think carefully about the behaviour of interest in relation to the information actually contained in our data. The work on domestic abuse in Chapter 2 highlights this instructively: We would like to have a record of all incidents of domestic abuse and which of those were reported to the police. Instead, we only have the reports to police which is a non-random subset of all incidents. Similarly in Chapter 1, we would like to know who was available or in sight of the police officer for each stop and search decision and why the officer ended up searching this particular individual. But even this idealized data would not contain the process by which the police officer ended up in this particular area. Further, even if we had our idealized stop and search data set, such data could not represent that people’s behaviour upon seeing a police officer might differ or that officers’ perception of suspicious behaviour depends on a person’s ethnicity (Alpert, MacDonald, and Dunham, 2005). This tension between the behaviour of interest and its data trace leads to several issues. We might draw false conclusions from spurious measurements, inappropriately interpret our results or misunderstand the dynamics of the behaviour of interest (Gelman and Loken, 2014; Hand, 2020). There is no immediate remedy to this tension, other than paying close attention and hopefully convincing more and more third parties to make their data and maybe even their data collection process available to researchers.

2. Model specification While it is a pre-condition to thoroughly understand the data-generating process, it is not sufficient and leads to the second aspect. More specifically, one needs the statistical toolkit to translate any hypotheses about the real world into mathematical models (Scheel et al., 2020). These models come with assumptions and conditions that need to at least approximately describe the true state of the world. If they do not, we might think ourselves safe from inappropriate inference. As I demonstrate in Chapter 3 however, even small violations of the assumption that the model is correctly specified drastically alters the inference. If the goal

of our research is the correct characterization of social phenomena, then proper inference is central to this endeavour. A key feature of this is engaging with bespoke models: For example, in the domestic abuse research in Chapter 2, it is not sufficient to realize the nature of the data-generating process and then use a standard model all the same with some minor adjustments. Instead, one needs to fundamentally rethink the appropriate statistical model and come to an approach that is informed by state of the art statistical methodology.

The solutions we develop to resolve these issues will always be context-dependent. In the domestic abuse example, we employ a point process to account for the continuous nature in which the realizations of the underlying phenomenon arrive. However for the stop and search work in Chapter 1, an entirely different inference approach is needed. There, we require a highly structured Bayesian hierarchical model so that we can situate our analysis at the appropriate (officer) level. But solving technical issues relating to data analysis is only part of the larger mission of doing (good) behavioural science.

I conclude the discussion of this thesis by addressing a further aspect: creating impactful work is central to behavioural science (Maner, 2016). Behavioural science can speak to processes influencing decisions that underlie many pressing policy issues, from ethnic disparities to inequality or climate change (Goff and Kahn, 2012; Lewandowsky, Cook, and Lloyd, 2018; Flèche, Lepinteur, and Powdthavee, 2020; Onyeador, Daumeyer, et al., 2020). Creating impact means engaging with the arguably most important stakeholders of research: policy makers.

Yet, this engagement with the policy sphere is difficult. While academics and policy makers may share a common goal of improving the state of the world, their time horizons, central questions, methods and knowledge differ significantly which leads to fundamental translation issues between the two (Fischhoff, 2019). For researchers, this means employing tools entirely different than those required for their academic work.

Part of this PhD project involved producing analyses for West Midlands Police. Additionally, I also participated in many stop and search commission meetings (public meetings of a mix of practitioners, police leadership as well as regular members of the public) to promote our research findings. I also joined an initiative

at West Midlands Police which aims to take a wider perspective on ethnic disparities by engaging academic researchers as well as police-internal and -external stakeholders. Translation issues on both sides create a gap between policy and academic expectations which needs to be bridged for research to succeed at its larger task of impact.

Policy makers are often used to thinking about a problem in terms of a specific quantity quantifying the problem. This problem is stark in the literature on ethnic disparities in stop and search. There are many different measures to quantify ethnic over-representation but in the United Kingdom, disproportionality ratios are the most commonly used one (e.g., Equality and Human Rights Commission, 2010). While there is some merit in the easy interpretation of a disproportionality ratio—a ratio of 4 means that Black people are searched at 4 times the rate of White people—the number of problems associated with using a ratio is large.

From a statistical point of view, ratios are very poorly behaved quantities (Pearson, 1897; Snedecor, 1946; Tanner, 1949; Neyman, 1952; Kronmal, 1993; Dunlap, Dietz, and Cortina, 1997) and very unstable in value when based on low counts. In the past, police forces have been reprimanded for disproportionate use of stop and search when these statistics were based on extraordinarily small numbers of searches (e.g., Equality and Human Rights Commission, 2010). The effectiveness of various interventions was then evaluated as a change in ratio value (Equality and Human Rights Commission, 2013). But typically, these ratios are presented without any form of uncertainty quantification and with small numbers these changes are easily random. Interventions were rated as (un)successful when not enough data was available to decide (Equality and Human Rights Commission, 2013). In spite of these limitations, disproportionality ratios remain the most commonly used quantity to make statements about stop and search in the United Kingdom. Changing such ingrained patterns of engaging statistics in the policy sphere is difficult but remains ultimately unavoidable if we want our research to have impact.

At the same time, there are often translation issues when it comes to the communication of research findings. Even simple probabilistic statements can create confusion. For example, I presented to stakeholders that annual stop and search rates for young Black men are 20%. Some of them then took this to mean that after five years of adulthood, a young Black men will have been searched with a probability of 100%, misunderstanding that the annual probabilities would multiply, not add up. This anecdote illustrates that academic researchers need to

think very carefully in how they present numerical quantities and how to communicate probabilities and uncertainty (Gigerenzer and Edwards, 2003; Gigerenzer, Gaissmaier, et al., 2007; Newall, 2016; Freeman, 2019; Blastland et al., 2020).

At the same time, researchers and policy makers alike may be motivated to gloss over the nuances of competing scientific studies in an effort to advance their own agendas (Lewandowsky, Mann, et al., 2016). Ethnic disparities in policing activities are a useful illustration of this difficulty: Persistent disparities in stop and search or use of force are sometimes presented as statistical artefacts due to imperfect analyses or data (e.g., The Centre for Social Justice, 2018; Johnson, Tress, et al., 2019). With unequal access to public attention, such research can gain prevalence in the public as academic diversity or even consensus, rather than a contradiction to a majority of studies documenting the opposite. Despite risks of institutional capture, closer collaboration between public institutions and academic research could potentially compel organisations to engage with the full range of academic research (Goff and Kahn, 2012).

Behavioural science has moved far in the last decade. The analysis of field data poses an exciting opportunity to promote our understanding of how people make decisions in the real world. At the same time, this development also poses new challenges and engaging with the data and model is only one component of the larger process of producing high quality science. In many ways, the replication crisis can be thought of as a self-correcting mechanism where scientists seek to improve the quality of the science they produce. The resulting increase in awareness and engagement with statistical methods in experimental settings will hopefully carry over into the analysis of field data. Many excellent tools exist to ensure that the findings behavioural scientists produce are generalisable, robust and impactful and I hope that my thesis could showcase some of them.

Bibliography

- Adebowale, V. (2013) *Independent Commission on Mental Health and Policing Report*. London: Independent Commission on Mental Health and Policing.
- Agüero, J. M. and Frisancho, V. (2021) Measuring Violence Against Women with Experimental Methods. *Economic Development and Cultural Change*, Forthcoming.
- Aizer, A. (2010) The Gender Wage Gap and Domestic Violence. *American Economic Review*, 100 (4), 1847–59.
- (2011) Poverty, violence, and health the impact of domestic violence during pregnancy on newborn health. *Journal of Human resources*, 46 (3), 518–538.
- Aizer, A. and Currie, J. (2004) Networks or neighborhoods? Correlations in the use of publicly-funded maternity care in California. *Journal of Public Economics*, 88 (12), 2573–2585.
- Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., et al. (2019) Preregistration of secondary data analysis: A template and tutorial. [Preprint] Available from: <https://psyarxiv.com/hvfmr>.
- Alevy, J. E., Haigh, M. S., and List, J. A. (2007) Information cascades: Evidence from a field experiment with financial market professionals. *The Journal of Finance*, 62 (1), 151–180.
- Alm, J., Bloomquist, K. M., and McKee, M. (2015) On the external validity of laboratory tax compliance experiments. *Economic Inquiry*, 53 (2), 1170–1186.
- Alpert, G. P., MacDonald, J. M., and Dunham, R. G. (2005) Police suspicion and discretionary decision making during citizen stops. *Criminology*, 43 (2), 407–434.
- Alpert, G. P., Smith, M. R., and Dunham, R. G. (2004) Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Research and Policy*, 6 (1), 43–69.

- American Civil Liberties Union (2009) *The Persistence of Racial and Ethnic Profiling in the United States*. New York: Rights Working Group.
- Anselin, L. (1988) *Spatial econometrics: methods and models*. Dordrecht: Springer Science & Business Media.
- Antonovics, K. and Knight, B. G. (2009) A new look at racial profiling: Evidence from the Boston Police Department. *The Review of Economics and Statistics*, 91 (1), 163–177.
- Aronow, P. M., Eckles, D., Samii, C., and Zonszein, S. (2020) Spillover Effects in Experimental Data. [Preprint] Available from: <https://arxiv.org/abs/2001.05444>.
- Ayres, I. (2002) Outcome tests of racial disparities in police practices. *Justice Research and Policy*, 4 (1-2), 131–142.
- Bachman, R. and Coker, A. L. (1995) Police involvement in domestic violence: The interactive effects of victim injury, offender’s history of violence, and race. *Violence and Victims*, 10 (2), 91–106.
- Baddeley, A., Turner, R., Møller, J., and Hazelton, M. (2005) Residual analysis for spatial point processes (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67 (5), 617–666.
- Barberis, N. C. (2013) Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives*, 27 (1), 173–96.
- Barnes, G. (2019) Lifetime Traffic Penalty Comparisons. *Freedom of Information Request 2019 1912 by the Guardian Australia*, Available from: <https://www.scribd.com/document/445453320/Document-for-Release-the-Guardian-FOI-2019-1912>. [Accessed 2021-02-09].
- Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019) Minimum Stein Discrepancy Estimators. *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 12964–12976.
- Bartlett, T. (2019) The Criminologist Accused of Cooking the Books. *The Chronicle of Higher Education*, [Online] Available from: <https://www.chronicle.com/article/the-criminologist-accused-of-cooking-the-books/>. [Accessed 2021-02-08].
- Basu, A., Harris, I. R., Hjort, N. L., and Jones, M. C. (1998) Robust and Efficient Estimation by Minimising a Density Power Divergence. *Biometrika*, 85 (3), 549–559.

- Bayer, P., Hjalmarsson, R., and Pozen, D. (2009) Building criminal capital behind bars: Peer effects in juvenile corrections. *The Quarterly Journal of Economics*, 124 (1), 105–147.
- Becker, G. S. (1957) *The economics of discrimination*. Chicago: University of Chicago Press.
- Belot, M., Duch, R., and Miller, L. (2015) A comprehensive comparison of students and non-students in classic experimental games. *Journal of Economic Behavior & Organization*, 113, 26–33.
- Bentham, M. (2021) Met chief: We will continue ‘disproportionate’ stop-and-search. *Evening Standard*, [Online] Available from: <https://www.standard.co.uk/news/uk/met-police-london-stop-and-search-racism-b918169.html>. [Accessed 2021-02-03].
- Beran, R. (1977) Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics*, 5 (3), 445–463.
- Bernasco, W. (2008) Them again? Same-offender involvement in repeat and near repeat burglaries. *European Journal of Criminology*, 5 (4), 411–431.
- Bernasco, W., Johnson, S. D., and Ruiter, S. (2015) Learning where to offend: Effects of past on future burglary locations. *Applied Geography*, 60, 120–129.
- Bertrand, M., Luttmer, E. F., and Mullainathan, S. (2000) Network effects and welfare cultures. *The Quarterly Journal of Economics*, 115 (3), 1019–1055.
- Beyer, K., Wallis, A. B., and Hamberger, L. K. (2015) Neighborhood environment and intimate partner violence: A systematic review. *Trauma, Violence, & Abuse*, 16 (1), 16–47.
- Biaggio, M. K., Brownell, A., and Watts, D. L. (1991) Reporting and seeking support by victims of sexual offenses. *Journal of Offender Rehabilitation*, 17 (1–2), 33–42.
- Bissiri, P. G., Holmes, C., and Walker, S. (2016) A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78 (5), 1103–1130.
- Black, V. (1955) Laboratory versus field research in psychology and the social sciences. *The British Journal for the Philosophy of Science*, 5 (20), 319–330.
- Blastland, M., Freeman, A. L., Linden, S. van der, Marteau, T. M., and Spiegelhalter, D. (2020) Five rules for evidence communication. *Nature*, 587.
- Bobonis, G. J., González-Brenes, M., and Castro, R. (2013) Public Transfers and Domestic Violence: The Roles of Private Information and Spousal Control. *American Economic Journal: Economic Policy*, 5 (1), 179–205.

- Bohren, A., Imas, A., and Rosenberg, M. (2018) The Language of Discrimination: Using Experimental versus Observational Data. *AEA Papers and Proceedings*, 108, 169–74.
- Bowling, B. and Phillips, C. (2007) Disproportionate and discriminatory: Reviewing the evidence on police stop and search. *The Modern Law Review*, 70 (6), 936–961.
- Bowsher, K. (2007) *The code systems used within the Metropolitan Police Service (MPS) to formally record ethnicity*. Tech. rep. 03/07. Metropolitan Police Service.
- Bradford, B. (2015) Unintended consequences. *Stop and Search: The Anatomy of a Police Power*. Ed. by M. Delsol Rebekah and Shiner. Basingstoke: Palgrave Macmillan, 102–122.
- (2017) *Stop and search and police legitimacy*. London: Routledge.
- Brantingham, P. J., Yuan, B., Sundback, N., Schoenberg, F. P., Bertozzi, A. L., Gordon, J., et al. (2018) Does violence interruption work?. *Proceedings of the National Academy of Sciences*, 8 (7), 1–6.
- Brayne, S. (2017) Big data surveillance: The case of policing. *American Sociological Review*, 82 (5), 977–1008.
- Brimicombe, A. and Cafe, R. (2012) Beware, win or lose: Domestic violence and the World Cup. *Significance*, 9 (5), 32–35.
- Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019) Statistical Inference for Generative Models with Maximum Mean Discrepancy. [Preprint] Available from: <https://arxiv.org/abs/1906.05944>.
- Brown, E. N., Barbieri, R., Ventura, V., Kass, R. E., and Frank, L. M. (2002) The time-rescaling theorem and its application to neural spike train data analysis. *Neural Computation*, 14 (2), 325–346.
- Bryan, G., Grant, M., Haggag, K., Karlan, D., Startz, M., and Udry, C. (2020) Blue Porches: Finding the limits of external validity of the endowment effect. *Journal of Economic Behavior & Organization*, 176, 269–271.
- Buyalskaya, A., Gallo, M., and Camerer, C. F. (2021) The golden age of social science. *Proceedings of the National Academy of Sciences*, 118 (5).
- Cai, J., De Janvry, A., and Sadoulet, E. (2015) Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7 (2), 81–108.
- Cameron, A. C. and Trivedi, P. K. (2013) *Regression analysis of count data*. Cambridge: Cambridge University Press.

- Card, D. and Dahl, G. B. (2011) Family violence and football: The effect of unexpected emotional cues on violent behavior. *The Quarterly Journal of Economics*, 126 (1), 103–143.
- Carlson, B. E. (2000) Children exposed to intimate partner violence: Research findings and implications for intervention. *Trauma, Violence, & Abuse*, 1 (4), 321–342.
- Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Goodrich, B., Betancourt, M., et al. (2017) Stan: A Probabilistic Programming Language. *Journal of Statistical Software*, 76 (1), 1–32.
- Casals, M., Girabent-Farres, M., and Carrasco, J. L. (2014) Methodological quality and reporting of generalized linear mixed models in clinical medicine (2000–2012): a systematic review. *PLoS One*, 9 (11), e112653.
- Centre for Women’s Justice (2020) Failure to address police perpetrated domestic abuse. Super-complaint by Centre for Women’s Justice. Available from: <https://www.centreforwomensjustice.org.uk/news/2020/3/9/police-officers-allowed-to-abuse-with-impunity-in-the-locker-room-culture-of-uk-forces-super-complaint-reveals>. [Accessed 2020-12-28].
- Cesario, J., Johnson, D. J., and Terrill, W. (2019) Is there evidence of racial disparity in police use of deadly force? Analyses of officer-involved fatal shootings in 2015–2016. *Social Psychological and Personality Science*, 10 (5), 586–595.
- Charles, K. K. and Guryan, J. (2011) Studying discrimination: Fundamental challenges and recent progress. *Annual Review of Economics*, 3 (1), 479–511.
- Charness, G. and Villeval, M.-C. (2009) Cooperation and competition in intergenerational experiments in the field and the laboratory. *American Economic Review*, 99 (3), 956–78.
- Cheng, I.-H. and Hsiaw, A. (2020) Reporting sexual misconduct in the #MeToo era. *American Economic Journal: Microeconomics*, [Forthcoming] Available from: http://people.brandeis.edu/~ahsiaw/chenghsiaw_reportingmisconduct.pdf. [Accessed 2020-12-10].
- Chérif-Abdellatif, B.-E. and Alquier, P. (2020a) Finite sample properties of parametric MMD estimation: robustness to misspecification and dependence. [Preprint] Available from: <https://arxiv.org/abs/1912.05737>.
- (2020b) MMD-Bayes: Robust Bayesian estimation via Maximum Mean discrepancy. *Symposium on Advances in Approximate Bayesian Inference*, 1–21.
- Clarke, R. V. and Cornish, D. B. (1985) Modeling offenders’ decisions: A framework for research and policy. *Crime and justice*, 6, 147–185.

- Cohen, L. E. and Felson, M. (1979) Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44 (4), 588–608.
- Cohn, E. G. (1993) The prediction of police calls for service: The influence of weather and temporal variables on rape and domestic violence. *Journal of Environmental Psychology*, 13 (1), 71–83.
- Committee on the Elimination of Racial Discrimination (2015) *Concluding observations on the combined nineteenth to twenty-second periodic reports of Germany*. International Convention on the Elimination of All Forms of Racial Discrimination CERD/C/DEU/CO/19-22.
- Cordy, C. B. and Griffith, D. A. (1993) Efficiency of least squares estimators in the presence of spatial autocorrelation. *Communications in Statistics-Simulation and Computation*, 22 (4), 1161–1179.
- Correll, J., Park, B., Judd, C. M., Wittenbrink, B., Sadler, M. S., and Keesee, T. (2007) Across the thin blue line: police officers and racial bias in the decision to shoot. *Journal of Personality and Social Psychology*, 92 (6), 1006–1023.
- Coy, M. and Kelly, L. (2011) *Islands in the Stream: An Evaluation of Four Independent Domestic Violence Advocacy Schemes*. Available from: <https://cwasu.org/wp-content/uploads/2016/07/IDVA-Main-Report1.pdf>. [Accessed 2020-10-02]. London: Henry Smith Charity, London Metropolitan University and Trust for London.
- Cramer, H. and Carter, M. (2002) *Homelessness: what's gender got to do with it*. London: Shelter.
- Daley, D. J. and Vere-Jones, D. (2003) *An introduction to the theory of point processes, volume 1: Elementary theory and methods*. New York: Springer.
- Davidson, J. (1994) *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford: Oxford University Press.
- Davis, R. C., Weisburd, D., and Hamilton, E. E. (2010) Preventing repeat incidents of family violence: A randomized field test of a second responder program. *Journal of Experimental Criminology*, 6 (4), 397–418.
- Davis, R. C., Weisburd, D., and Taylor, B. (2008) Effects of second responder programs on repeat incidents of family abuse: A systematic review. *Campbell Systematic Reviews*, 4 (1), 1–38.
- Del Toro, J., Lloyd, T., Buchanan, K. S., Robins, S. J., Bencharit, L. Z., Smiedt, M. G., et al. (2019) The criminogenic and psychological effects of police stops on adolescent black and Latino boys. *Proceedings of the National Academy of Sciences*, 116 (17), 8261–8268.

- DellaVigna, S. (2009) Psychology and economics: Evidence from the field. *Journal of Economic literature*, 47 (2), 315–72.
- Delsol, R. (2015) Effectiveness. *Stop and Search: The Anatomy of a Police Power*. Ed. by M. Delsol Rebekah and Shiner. Basingstoke: Palgrave Macmillan, 79–101.
- Delsol, R. and Shiner, M. (2006) Regulating stop and search: A challenge for police and community relations in England and Wales. *Critical Criminology*, 14 (3), 241–263.
- Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L1 View*. New York: Wiley.
- Devroye, L. and Lugosi, G. (2012) *Combinatorial methods in density estimation*. New York: Springer Science & Business Media.
- Dharmapala, D. and Ross, S. L. (2004) Racial bias in motor vehicle searches: Additional theory and evidence. *The B.E. Journal of Economic Analysis & Policy*, 3 (1), 1–23.
- Van Dolder, D., van den Assem, M. J., Camerer, C. F., and Thaler, R. H. (2015) Standing United or Falling Divided? High Stakes Bargaining in a TV Game Show. *American Economic Review*, 105 (5), 402–407.
- Drago, F., Mengel, F., and Traxler, C. (2020) Compliance behavior in networks: Evidence from a field experiment. *American Economic Journal: Applied Economics*, 12 (2), 96–133.
- Dunlap, W. P., Dietz, J., and Cortina, J. M. (1997) The spurious correlation of ratios that have common variables: A Monte Carlo examination of Pearson’s formula. *The Journal of General Psychology*, 124 (2), 182–193.
- Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., et al. (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS One*, 3 (8), e3081.
- Eberhardt, J. L., Goff, P. A., Purdie, V. J., and Davies, P. G. (2004) Seeing black: Race, crime, and visual processing. *Journal of Personality and Social Psychology*, 87 (6), 876–893.
- Eckhouse, L. (2018) Everyday Risk–Disparate Exposure and Racial Inequality in Police Violence. [Preprint] Available from: bit.ly/FxqHJD. [Accessed 2021-02-09].
- Edwards, F., Esposito, M. H., and Lee, H. (2018) Risk of police-involved death by race/ethnicity and place, United States, 2012–2018. *American Journal of Public Health*, 108 (9), 1241–1248.

- Efron, B. (1979) Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 7 (1), 1–26.
- Elliott, D. S. et al. (1995) *Lies, damn lies, and arrest statistics*. Boulder, CO: Center for the Study and Prevention of Violence.
- Ellsberg, M., Arango, D. J., Morton, M., Gennari, F., Kiplesund, S., Contreras, M., et al. (2015) Prevention of violence against women and girls: what does the evidence say?. *The Lancet*, 385 (9977), 1555–1566.
- Ellsberg, M., Heise, L., Pena, R., Agurto, S., and Winkvist, A. (2001) Researching domestic violence against women: methodological and ethical considerations. *Studies in Family Planning*, 32 (1), 1–16.
- Emery, C. R. (2010) Examining an extension of Johnson’s hypothesis: Is male perpetrated intimate partner violence more underreported than female violence?. *Journal of Family Violence*, 25 (2), 173–181.
- Engel, R. S. and Tillyer, R. (2008) Searching for equilibrium: The tenuous nature of the outcome test. *Justice Quarterly*, 25 (1), 54–71.
- Engel, R. S., Sobol, J. J., and Worden, R. E. (2000) Further exploration of the demeanor hypothesis: The interaction effects of suspects’ characteristics and demeanor on police behavior. *Justice Quarterly*, 17 (2), 235–258.
- Equality and Human Rights Commission (2010) *Stop and think: A critical review of the use of stop and search powers in England and Wales*. London: Equality and Human Rights Commission.
- (2013) *Stop and think again, towards equality in police PACE stop and search*. London: Equality and Human Rights Commission.
- Erwin, P. E. (2006) Exporting US domestic violence reforms: An analysis of human rights frameworks and US “best practices”. *Feminist Criminology*, 1 (3), 188–206.
- Fadlon, I. and Nielsen, T. H. (2019) Family health behaviors. *American Economic Review*, 109 (9), 3162–3191.
- Fagan, J. (1996) *The criminalization of domestic violence: Promises and limits*. National Institute of Justice Research Report. US Department of Justice.
- Fagan, J. and Davies, G. (2000) Street stops and broken windows: Terry, race, and disorder in New York City. *Fordham Urban Law Journal*, 28 (2), 457–504.
- Fanelli, D. (2010) “Positive” results increase down the hierarchy of the sciences. *PLoS One*, 5 (4), e10068.
- Farrell, G. and Pease, K. (1993) *Once bitten, twice bitten: Repeat victimisation and its implications for crime prevention*. Crime Prevention Unit Paper 46. Home Office Police Research Group London.

- Fatsis, L. (2019) Policing the beats: The criminalisation of UK drill and grime music by the London Metropolitan Police. *The Sociological Review*, 67 (6), 1300–1316.
- Felson, R. B., Messner, S. F., Hoskin, A. W., and Deane, G. (2002) Reasons for reporting and not reporting domestic violence to the police. *Criminology*, 40 (3), 617–648.
- Fischhoff, B. (2019) Evaluating science communication. *Proceedings of the National Academy of Sciences*, 116 (16), 7670–7675.
- Fisher, B. S., Daigle, L. E., Cullen, F. T., and Turner, M. G. (2003) Reporting sexual victimization to the police and others: Results from a national-level study of college women. *Criminal Justice and Behavior*, 30 (1), 6–38.
- Fitzgerald, L. F., Swan, S., and Fischer, K. (1995) Why didn't she just report him? The psychological and legal implications of women's responses to sexual harassment. *Journal of Social Issues*, 51 (1), 117–138.
- Flaxman, S., Chirico, M., Pereira, P., Loeffler, C., et al. (2019) Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: A winning solution to the NIJ “Real-Time Crime Forecasting Challenge”. *The Annals of Applied Statistics*, 13 (4), 2564–2585.
- Flèche, S., Lepinteur, A., and Powdthavee, N. (2020) Gender norms, fairness and relative working hours within households. *Labour Economics*, 65, OnlineFirst. DOI: <https://doi.org/10.1016/j.labeco.2020.101866>.
- Folke, O. and Rickne, J. K. (2020) *Sexual Harassment and Gender Inequality in the Labor Market*. Discussion Paper 14737. Centre for Economic Policy Research.
- Fong, E., Lyddon, S., and Holmes, C. (2019) Scalable Nonparametric Sampling from Multimodal Posteriors with the Posterior Bootstrap. [Preprint] Available from: <https://arxiv.org/abs/1902.03175>.
- Fox, C. R., Erner, C., and Walters, D. J. (2015) Decision under risk: From the field to the lab and back. *The Wiley Blackwell Handbook of Judgment and Decision Making*. Ed. by G. Keren and G. Wu. Hoboken: John Wiley & Sons, Ltd, 41–88.
- Franco, A., Malhotra, N., and Simonovits, G. (2014) Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345 (6203), 1502–1505.
- Freedman, D. A. (1999) Ecological inference and the ecological fallacy. *International Encyclopedia of the Social & Behavioral sciences*. Ed. by N. Smelser and P. Baltes. Vol. 6. New York: Elsevier, 4027–4030.

- Freeman, A. L. (2019) How to communicate evidence to patients. *Drug and Therapeutics Bulletin*, 57 (8), 119–124.
- Freisthler, B. and Weiss, R. E. (2008) Using Bayesian space-time models to understand the substance use environment and risk for being referred to child protective services. *Substance Use & Misuse*, 43 (2), 239–251.
- Fryer, R. G. (2019) An empirical analysis of racial differences in police use of force. *Journal of Political Economy*, 127 (3), 1210–1261.
- Futami, F., Sato, I., and Sugiyama, M. (2018) Variational Inference based on Robust Divergences. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by A. Storkey and F. Perez-Cruz. Vol. 84. Cambridge: Proceedings of Machine Learning Research, 813–822.
- Galizzi, M. M. and Navarro-Martinez, D. (2019) On the external validity of social preference games: a systematic lab-field study. *Management Science*, 65 (3), 976–1002.
- Galizzi, M. M. and Whitmarsh, L. (2019) How to measure behavioral spillovers: a methodological review and checklist. *Frontiers in Psychology*, 10, 342.
- Gao, K. and Khoshgoftaar, T. M. (2007) A comprehensive empirical study of count models for software fault prediction. *IEEE Transactions on Reliability*, 56 (2), 223–236.
- Gartner, R. and Macmillan, R. (1995) The effect of victim-offender relationship on reporting crimes of violence against women. *Canadian Journal of Criminology*, 37 (3), 393–429.
- Gehlke, C. E. and Biehl, K. (1934) Certain effects of grouping upon the size of the correlation coefficient in census tract material. *Journal of the American Statistical Association*, 29 (185A), 169–170.
- Geller, A., Fagan, J., Tyler, T., and Link, B. G. (2014) Aggressive policing and the mental health of young urban men. *American Journal of Public Health*, 104 (12), 2321–2327.
- Gelman, A., Fagan, J., and Kiss, A. (2007) An analysis of the New York City Police department’s ‘stop-and-frisk’ policy in the context of claims of racial bias. *Journal of the American Statistical Association*, 102 (479), 813–823.
- Gelman, A. and Loken, E. (2014) The statistical crisis in science. *American Scientist*, 102 (6), 460–465.
- Georgii, H.-O. (1976) Canonical and grand canonical Gibbs states for continuum systems. *Communications in Mathematical Physics*, 48 (1), 31–51.
- Germain, P., Bach, F., Lacoste, A., and Lacoste-Julien, S. (2016) PAC-Bayesian theory meets Bayesian inference. *Advances in Neural Information Processing*

- Systems*. Ed. by D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett. Vol. 29. Curran Associates, Inc., 1884–1892.
- Ghosal, S. (1996) A review of consistency and convergence rates of posterior distributions. *Proceedings of Varanashi Symposium in Bayesian Inference*.
- Ghosal, S. and Van der Vaart, A. (2007) Convergence rates of posterior distributions for non-iid observations. *The Annals of Statistics*, 35 (1), 192–223.
- Ghosh, A. and Basu, A. (2016) Robust Bayes estimation using the density power divergence. *Annals of the Institute of Statistical Mathematics*, 68 (2), 413–437.
- Gigerenzer, G. (2004) Mindless statistics. *The Journal of Socio-Economics*, 33 (5), 587–606.
- Gigerenzer, G. and Edwards, A. (2003) Simple tools for understanding risks: from innumeracy to insight. *British Medical Journal*, 327, 741–744.
- Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., and Woloshin, S. (2007) Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, 8 (2), 53–96.
- Goel, S., Rao, J. M., and Shroff, R. (2016) Precinct or prejudice? Understanding racial disparities in New York City’s stop-and-frisk policy. *The Annals of Applied Statistics*, 10 (1), 365–394.
- Goff, P. A. and Kahn, K. B. (2012) Racial bias in policing: Why we know less than we should. *Social Issues and Policy Review*, 6 (1), 177–210.
- Goff, P. A., Steele, C. M., and Davies, P. G. (2008) The space between us: stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology*, 94 (1), 91–107.
- Goh, J. X., Bandt-Law, B., Cheek, N. N., Sinclair, S., and Kaiser, C. R. (2021) Narrow prototypes and neglected victims: Understanding perceptions of sexual harassment. *Journal of Personality and Social Psychology*, Online ahead of print.
- Goncalves, F. and Mello, S. (2020) A few bad apples? Racial bias in policing. [Preprint] Available from: <http://dx.doi.org/10.2139/ssrn.3627809>. [Accessed 2021-02-05].
- Goodkind, J. R., Gillum, T. L., Bybee, D. I., and Sullivan, C. M. (2003) The impact of family and friends’ reactions on the well-being of women with abusive partners. *Violence Against Women*, 9 (3), 347–373.
- Gorr, W. L. and Lee, Y. (2015) Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology*, 31 (1), 25–47.

- Gottfredson, M. R. and Gottfredson, D. M. (1988) *Decision making in criminal justice: Toward the rational exercise of discretion*. New York: Springer Science & Business Media.
- Goudriaan, H., Wittebrood, K., and Nieuwbeerta, P. (2006) Neighbourhood Characteristics and Reporting Crime: Effects of Social Cohesion, Confidence in Police Effectiveness and Socio-Economic Disadvantage. *British Journal of Criminology*, 46 (4), 719–742.
- Gounev, P. and Bezlov, T. (2006) The Roma in Bulgaria’s criminal justice system: From ethnic profiling to imprisonment. *Critical Criminology*, 14 (3), 313–338.
- Gracia, E., López-Quílez, A., Marco, M., Lladosa, S., and Lila, M. (2015) The spatial epidemiology of intimate partner violence: do neighborhoods matter?. *American Journal of Epidemiology*, 182 (1), 58–66.
- Greenberg, M. S. and Ruback, R. B. (1992) *After the crime: Victim decision making*. New York: Springer Science & Business Media.
- Greenfeld, L. A., Rand, M. R., Craven, D., Klaus, P. A., Perkins, C. A., Ringel, C., et al. (1998) *Violence by intimates: Analysis of data on crimes by current or former spouses, boyfriends, and girlfriends*. Washington, DC: U.S. Department of Justice–Bureau of Justice Statistics.
- Guedj, B. (2019) A primer on PAC-Bayesian learning. [Preprint] Available from: <https://arxiv.org/abs/1901.05353>.
- Haining, R. and Law, J. (2007) Combining police perceptions with police records of serious crime areas: a modelling approach. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 170 (4), 1019–1034.
- Hall, P. and Turlach, B. A. (1999) Reducing bias in curve estimation by use of weights. *Computational Statistics & Data Analysis*, 30 (1), 67–86.
- Halloran, M. E. and Hudgens, M. G. (2016) Dependent happenings: a recent methodological review. *Current Epidemiology Reports*, 3 (4), 297–305.
- Hand, D. J. (2020) *Dark Data: Why What You Don’t Know Matters*. Princeton: Princeton University Press.
- Hanmer, J., Griffiths, S., and Jerwood, D. (1999) *Arresting evidence: domestic violence and repeat victimisation*. Police Research Series Paper 104. Home Office, Policing & Reducing Crime Unit.
- Harrison, G. W., Lau, M. I., and Williams, M. B. (2002) Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, 92 (5), 1606–1617.

- Hawinkel, S., Rayner, J. C. W., Bijmens, L., and Thas, O. (2020) Sequence count data are poorly fit by the negative binomial distribution. *PLoS One*, 15 (4), e0224909.
- Hawkes, A. G. (1971) Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58 (1), 83–90.
- Hawkins, S. and Laxton, C. (2014) *Women’s access to justice: from reporting to sentencing*. London: All-Party Parliamentary Group on Domestic and Sexual Violence.
- Heckman, J. J. (1991) Identifying the hand of past: Distinguishing state dependence from heterogeneity. *The American Economic Review*, 81 (2), 75–79.
- Heise, L. L. (1998) Violence against women: An integrated, ecological framework. *Violence Against Women*, 4 (3), 262–290.
- Henrich, J., Heine, S. J., and Norenzayan, A. (2010) The weirdest people in the world?. *Behavioral and Brain Sciences*, 33 (2-3), 61–83.
- Her Majesty’s Inspectorate of Constabulary (2013) *Stop and search powers: Are the police using them effectively and fairly?* London: Her Majesty’s Inspectorate of Constabulary.
- (2014) *Everyone’s business: Improving the police response to domestic abuse*. London: Her Majesty’s Inspectorate of Constabulary.
- Hester, M. (2006) Making it through the criminal justice system: Attrition and domestic violence. *Social Policy and Society*, 5 (1), 79.
- Ver Hoef, J. M. and Boveng, P. L. (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. *Ecology*, 88 (11), 2766–2772.
- Holroyd, J. and Sweetman, J. (2016) The heterogeneity of Implicit Bias. *Philosophy and Implicit Bias: Volume 1—Metaphysics and Epistemology*. Ed. by J. Saul and M. Brownstein. Oxford: Oxford University Press, 80–103.
- Home Office (2013) *New government domestic violence and abuse definition*. Correspondence: Circular 003/2013. Home Office.
- (2014) *CODE A—Revised Code of Practice for the exercise by: Police Officers of Statutory Powers of stop and search*. Norwich: Her Majesty’s Stationary Office.
- (2017) Table A.01c: Number of persons arrested by ethnic group year ending 31 March 2017. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/654461/arrest-police-powers-procedures-mar17-hosb2017-tables.ods.

- Home Office (2018a) *Police powers and procedures England and Wales year ending 31 March 2018*. London: Home Office Statistical Bulletin.
- (2018b) *Police powers and procedures, England and Wales, year ending 31 March 2018*. London: Home Office.
- Hong, G. and Raudenbush, S. W. (2006) Evaluating kindergarten retention policy: A case study of causal inference for multilevel observational data. *Journal of the American Statistical Association*, 101 (475), 901–910.
- Hooker, G. and Vidyashankar, A. N. (2014) Bayesian model robustness via disparities. *Test*, 23 (3), 556–584.
- Hoppe, S. J., Zhang, Y., Hayes, B. E., and Bills, M. A. (2020) Mandatory arrest for domestic violence and repeat offending: A meta-analysis. *Aggression and Violent Behavior*, 53, 101430.
- Human Rights Watch (2020) *“They Talk to Us Like We’re Dogs”—Abusive Police Stops in France*. New York: Human Rights Watch.
- Hyvärinen, A. (2005) Estimation of Non-Normalized Statistical Models by Score Matching. *Journal of Machine Learning Research*, 6, 695–708.
- Independent Office for Police Conduct (2018) *Investigation into Sussex Police contact with Shana Grice prior to her murder on 25 August 2016*. London: Independent Office for Police Conduct.
- Ivandic, R., Kirchmaier, T., and Linton, B. (2020) *Changing patterns of domestic abuse during Covid-19 lockdown*. Discussion Paper 1729. Centre for Economic Performance.
- Iyer, L., Mani, A., Mishra, P., and Topalova, P. (2012) The power of political voice: women’s political representation and crime in India. *American Economic Journal: Applied Economics*, 4 (4), 165–93.
- Jewkes, R. (2002) Intimate partner violence: causes and prevention. *The Lancet*, 359 (9315), 1423–1429.
- Jewson, J., Smith, J. Q., and Holmes, C. (2018) Principles of Bayesian inference using general divergence criteria. *Entropy*, 20 (6), 442.
- John, L. K., Loewenstein, G., and Prelec, D. (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23 (5), 524–532.
- Johnson, D. J., Tress, T., Burkel, N., Taylor, C., and Cesario, J. (2019) Officer characteristics and racial disparities in fatal officer-involved shootings. *Proceedings of the National Academy of Sciences*, 116 (32), 15877–15882.

- Johnson, S. D. and Bowers, K. J. (2004) The burglary as clue to the future: The beginnings of prospective hot-spotting. *European Journal of Criminology*, 1 (2), 237–255.
- Kalair, K., Connaughton, C., and Di Loro, P. A. (2020) A non-parametric Hawkes process model of primary and secondary accidents on a UK smart motorway. [Preprint] Available from: <https://arxiv.org/abs/2004.14194>.
- Kane, R. J. (2003) Social control in the metropolis: A community-level examination of the minority group-threat hypothesis. *Justice Quarterly*, 20 (2), 265–295.
- Kavanaugh, G., Sviatschi, M. M., and Trako, I. (2019) Female Officers, Gender Violence and Human Capital: Evidence from All-Women’s Justice Centers in Peru. [Preprint] Available from: <https://rpd.princeton.edu/sites/rpd/files/sviatschi-female-officers-gender-violence-and-children-feb2019.pdf>. [Accessed 2021-02-09].
- Kelly, L., Sharp-Jeffs, N., and Klein, R. (2014) *Finding the costs of freedom: How women and children rebuild their lives after domestic violence*. London: Solace.
- Kelly, L., Bindel, J., Burton, S., Butterworth, D., Cook, K., and Regan, L. (1999) *Domestic violence matters: An evaluation of a development project*. London: Home Office.
- Kelly, L. and Westmorland, N. (2016) Naming and defining ‘domestic violence’: Lessons from research with violent men. *Feminist Review*, 112 (1), 113–127.
- Kirby, S., Francis, B., and O’Flaherty, R. (2014) Can the FIFA world cup football (soccer) tournament be associated with an increase in domestic abuse?. *Journal of Research in Crime and Delinquency*, 51 (3), 259–276.
- Knoblauch, J. (2019) Frequentist Consistency of Generalized Variational Inference. [Preprint] Available from: <https://arxiv.org/abs/1912.04946>.
- Knoblauch, J., Jewson, J., and Damoulas, T. (2019) Generalized variational inference: Three arguments for deriving new posteriors. [Preprint] Available from: <https://arxiv.org/abs/1904.02063>.
- Knoblauch, J., Jewson, J. E., and Damoulas, T. (2018) Doubly Robust Bayesian Inference for Non-Stationary Streaming Data using β -Divergences. *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 64–75.
- Knowles, J., Persico, N., and Todd, P. (2001) Racial bias in motor vehicle searches: Theory and evidence. *Journal of Political Economy*, 109 (1), 203–229.

- Knox, D., Lowe, W., and Mummolo, J. (2020) Administrative Records Mask Racially Biased Policing. *American Political Science Review*, 114 (3), 619–637.
- Kohler-Hausmann, I. (2013) Misdemeanor justice: Control without conviction. *American Journal of Sociology*, 119 (2), 351–393.
- Kronmal, R. A. (1993) Spurious correlation and the fallacy of the ratio standard revisited. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 156 (3), 379–392.
- Kruschke, J. K. and Liddell, T. M. (2018) The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and power analysis from a Bayesian perspective. *Psychonomic Bulletin & Review*, 25 (1), 178–206.
- Kullback, S. and Leibler, R. A. (1951) On information and sufficiency. *The Annals of Mathematical Statistics*, 22 (1), 79–86.
- Lakens, D., Scheel, A. M., and Isager, P. M. (2018) Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science*, 1 (2), 259–269.
- Lambert, D. (1992) Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34 (1), 1–14.
- Lammy, D. (2017) *The Lammy review: An independent review into the treatment of, and outcomes for, Black, Asian and Minority Ethnic individuals in the Criminal Justice System*. London: Ministry of Justice.
- Laniyonu, A. (2018) Police, politics and participation: The effect of police exposure on political participation in the United Kingdom. *The British Journal of Criminology*, 58 (5), 1232–1253.
- Laub, J. H. (1981) Ecological considerations in victim reporting to the police. *Journal of Criminal Justice*, 9 (6), 419–430.
- Lea, J. (2000) The Macpherson Report and the question of institutional racism. *The Howard Journal of Criminal Justice*, 39 (3), 219–233.
- Legewie, J. and Fagan, J. (2016) *Group threat, police officer diversity and the deadly use of police force*. Columbia Public Law Research Paper 14-512. Columbia Law School.
- Leonard, K. (2001) Domestic violence and alcohol: what is known and what do we need to know to encourage environmental interventions?. *Journal of Substance Use*, 6 (4), 235–247.
- Leonard, K. E. and Quigley, B. M. (2017) Thirty years of research show alcohol to be a cause of intimate partner violence: Future research needs to identify who to treat and how to treat them. *Drug and Alcohol Review*, 36 (1), 7–9.

- Lerman, A. E. and Weaver, V. (2014) Staying out of sight? Concentrated policing and local political action. *The ANNALS of the American Academy of Political and Social Science*, 651 (1), 202–219.
- Levitt, S. D. and List, J. A. (2007a) On the generalizability of lab behaviour to the field. *Canadian Journal of Economics/Revue canadienne d'économique*, 40 (2), 347–370.
- (2007b) What do laboratory experiments measuring social preferences reveal about the real world?. *Journal of Economic Perspectives*, 21 (2), 153–174.
- Levy, R. and Mattsson, M. (2020) The effects of social movements: Evidence from #MeToo. [Preprint] Available from: <http://dx.doi.org/10.2139/ssrn.3496903>. [Accessed 2020-09-02].
- Lewandowsky, S., Cook, J., and Lloyd, E. (2018) The ‘Alice in Wonderland’ mechanics of the rejection of (climate) science: simulating coherence by conspiracism. *Synthese*, 195 (1), 175–196.
- Lewandowsky, S., Mann, M. E., Brown, N. J., and Friedman, H. (2016) Science and the public: Debate, denial, and skepticism. *Journal of Social and Political Psychology*, 4 (2), 537–553.
- Lindsey, J. K. and Jones, B. (1998) Choosing among generalized linear models applied to medical data. *Statistics in Medicine*, 17 (1), 59–68.
- Livingston, M. (2011) A longitudinal analysis of alcohol outlet density and domestic violence. *Addiction*, 106 (5), 919–925.
- Loeffler, C. and Flaxman, S. (2018) Is gun violence contagious? A spatiotemporal test. *Journal of Quantitative Criminology*, 34 (4), 999–1017.
- Lombard, N. and McMillan, L. (2013) *Violence against Women: Current Theory and Practice in Domestic Abuse, Sexual Violence and Exploitation*. London: Jessica Kingsley Publishers.
- Lombard, N. and Scott, M. (2013) Older women and domestic abuse: where ageism and sexism intersect. *Violence against women: Current theory and practice in domestic abuse, sexual violence and exploitation*. Ed. by N. Lombard and L. McMillan. London: Jessica Kingsley Publishers, 125–140.
- Lyddon, S., Walker, S., and Holmes, C. C. (2018) Nonparametric learning from Bayesian models with randomized objective functions. *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2071–2081.
- Lyddon, S., Holmes, C., and Walker, S. (2019) General Bayesian updating and the loss-likelihood bootstrap. *Biometrika*, 106 (2), 465–478.

- MacDonald, J., Fagan, J., and Geller, A. (2016) The effects of local police surges on crime and arrests in New York City. *PLoS One*, 11 (6), e0157223.
- Macpherson, W. (1999) *The Stephen Lawrence Inquiry*. London: Home Office.
- Malleson, N. and Andresen, M. A. (2015) The impact of using social media data in crime rate calculations: Shifting hot spots and changing spatial patterns. *Cartography and Geographic Information Science*, 42 (2), 112–121.
- Maner, J. K. (2016) Into the wild: Field research can increase both replicability and real-world impact. *Journal of Experimental Social Psychology*, 66, 100–106.
- Manousakas, D. and Mascolo, C. (2020) β -Cores: Robust Large-Scale Bayesian Data Summarization in the Presence of Outliers. [Preprint] Available from: <https://arxiv.org/abs/2008.13600>.
- Mardia, J., Jiao, J., Tánzos, E., Nowak, R. D., and Weissman, T. (2019) Concentration inequalities for the empirical distribution. [Preprint] Available from: <https://arxiv.org/abs/1809.06522>.
- Markham, F., Doran, B., and Young, M. (2016) The relationship between electronic gaming machine accessibility and police-recorded domestic violence: a spatio-temporal analysis of 654 postcodes in Victoria, Australia, 2005–2014. *Social Science & Medicine*, 162, 106–114.
- Mastrofski, S. D., Parks, R. B., and Worden, R. E. (1998) *Community policing in action: Lessons from an observational study*. Washington, DC: US Department of Justice.
- Matczak, A., Hatzidimitriadou, E., and Lindsay, J. (2011) *Review of domestic violence policies in England & Wales*. London: Kingston University and St George’s, University of London.
- Mawby, R. C. and Wright, A. (2008) The police organisation. *Handbook of Policing*. Ed. by T. Newburn. 2nd ed. Abingdon: Taylor & Francis, 224–228.
- May, T., Gyang, T., and Hough, M. (2010) *Differential treatment in the youth justice system*. London: Equality and Human Rights Commission.
- McCandless, R., Feist, A., Allan, J., and Morgan, N. (2016) *Do initiatives involving substantial increases in stop and search reduce crime? Assessing the impact of Operation Blunt 2*. London: Home Office.
- McCrary, J. (2007) The effect of court-ordered hiring quotas on the composition and quality of police. *American Economic Review*, 97 (1), 318–353.
- McDougal, L., Krumholz, S., Bhan, N., Bharadwaj, P., and Raj, A. (2018) Releasing the tide: how has a shock to the acceptability of gender-based sexual

- violence affected rape reporting to police in India?. *Journal of Interpersonal Violence*, OnlineFirst. DOI: <https://doi.org/10.1177/0886260518811421>.
- Merriam-Webster (2021) *Behavioral science*. *Merriam-Webster.com dictionary*. [Online] Available from: <https://www.merriam-webster.com/dictionary/behavioral%20science>. [Accessed 2021-02-04].
- Meyer, S., Warnke, I., Rössler, W., and Held, L. (2016) Model-based testing for space-time interaction using point processes: An application to psychiatric hospital admissions in an urban area. *Spatial and Spatio-temporal Epidemiology*, 17, 15–25.
- Mickish, J. E. (2002) Domestic Violence. *Crisis Intervention in Criminal Justice/Social Service*. Ed. by B. D. Byers and J. E. Hendricks. Springfield, 77–118.
- Miller, A. R. and Segal, C. (2019) Do female officers improve law enforcement quality? Effects on crime reporting and domestic violence. *The Review of Economic Studies*, 86 (5), 2220–2247.
- Miller, J. (2019) Asymptotic normality, concentration, and coverage of generalized posteriors. [Preprint] Available from: <https://arxiv.org/abs/1907.09611>.
- Miller, J., Le Masurier, P., and Wicks, J. (2000) *Profiling populations available for stops and searches*. Police Research Series Paper 131. Home Office, Policing & Reducing Crime Unit.
- Mitchell, G. (2012) Revisiting truth or triviality: The external validity of research in the psychological laboratory. *Perspectives on Psychological Science*, 7 (2), 109–117.
- Mohler, G. (2014) Marked point process hotspot maps for homicide and gun crime prediction in Chicago. *International Journal of Forecasting*, 30 (3), 491–497.
- Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., and Tita, G. E. (2011) Self-exciting point process modeling of crime. *Journal of the American Statistical Association*, 106 (493), 100–108.
- Mook, D. (1983) In Defense of External Invalidity. *American Psychologist*, 38 (4), 379–387.
- Morris, W., Burden, A., and Weekes, A. (2004) *The Report of the Morris Inquiry—The case for change: People in the Metropolitan Police Service*. London: Metropolitan Police Authority.
- Muchow, A. N. and Amuedo-Dorantes, C. (2020) Immigration enforcement awareness and community engagement with police: Evidence from domestic violence calls in Los Angeles. *Journal of Urban Economics*, 117, 103253.

- Myhill, A. and Allen, J. (2002) *Rape and sexual assault of women: the extent and nature of the problem*. Home Office Research Study 237. Home Office Research, Development and Statistics Directorate.
- Neil, R. and Winship, C. (2018) Methodological challenges and opportunities in testing for racial discrimination in policing. *Annual Review of Criminology*, 2, 73–98.
- Nelson, E. L. (2013) Police controlled antecedents which significantly elevate prosecution and conviction rates in domestic violence cases. *Criminology & Criminal Justice*, 13 (5), 526–551.
- Newall, P. W. (2016) Downside financial risk is misunderstood. *Judgement and Decision Making*, 11 (5), 416–423.
- Neyman, J. (1952) *Lectures and conferences on mathematical statistics and probability*. Washington, DC: Graduate School, US Department of Agriculture.
- Nguyen, X. X. and Zessin, H. (1979) Integral and differential characterizations of the Gibbs process. *Mathematische Nachrichten*, 88 (1), 105–115.
- Nickerson, D. W. (2008) Is voting contagious? Evidence from two field experiments. *The American Political Science Review*, 102 (1), 49–57.
- Nicoletti, C., Salvanes, K. G., and Tominey, E. (2018) The family peer effect on mothers’ labor supply. *American Economic Journal: Applied Economics*, 10 (3), 206–34.
- Nixon, J. (2009) Domestic violence and women with disabilities: locating the issue on the periphery of social movements. *Disability & Society*, 24 (1), 77–89.
- O’Neal, E. N. (2019) “Victim is not credible”: The influence of rape culture on police perceptions of sexual assault complainants. *Justice Quarterly*, 36 (1), 127–160.
- Oberfield, Z. W. (2012) Socialization and self-selection: How police officers develop their views about using force. *Administration & Society*, 44 (6), 702–730.
- Office for National Statistics (2003) *Ethnic group statistics—A guide for the collection and classification of ethnicity data*. London: Office for National Statistics.
- (2009) *Final recommended questions for the 2011 Census in England and Wales—National Identity*. London: Office for National Statistics.
- (2011a) 2011 Census: QS201EW Ethnic group: countries. Available from: <https://www.nomisweb.co.uk/census/2011/qs211ew>. [Accessed 2020-12-07].

- Office for National Statistics (2011b) Table KS401EW - Dwellings, household spaces and accommodation type by 2011 output area. Available from: <https://www.nomisweb.co.uk/census/2011/ks401ew>. [Accessed 2020-12-15].
- (2015) National Statistics—English indices of deprivation 2015 File 1. Available from: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/467764/File_1_ID_2015_Index_of_Multiple_Deprivation.xlsx. [Accessed 2020-12-15].
- (2016) Census geography—An overview of the various geographies used in the production of statistics collected via the UK census. Available from: <https://web.archive.org/web/20190715091529/https://www.ons.gov.uk/methodology/geography/ukgeographies/censusgeography>. [Accessed 2019-07-15].
- (2017) *Crime Survey for England and Wales Technical Report 2018/19—Volume 1*. Available from: <https://www.ons.gov.uk/file?uri=/peoplepopulationandcommunity/crimeandjustice/methodologies/crimeandjusticemethodology/201819csewtechnicalpdf>. [Accessed 2021-02-10]. London: Kantar.
- (2019a) Domestic abuse prevalence and trends, England and Wales: year ending March 2019. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/domesticabuseprevalenceandtrendsenglandandwales/yearendingmarch2019>. [Accessed 2021-02-10].
- (2019b) UK homelessness: 2005 to 2018. Available from: <https://www.ons.gov.uk/peoplepopulationandcommunity/housing/articles/ukhomelessness/2005to2018#reasons-for-homelessness>. [Accessed 2021-02-09].
- Ogata, Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 83 (401), 9–27.
- Onyeador, I. N., Daumeyer, N. M., Rucker, J. M., Duker, A., Kraus, M. W., and Richeson, J. A. (2020) Disrupting beliefs in racial progress: Reminders of persistent racism alter perceptions of past, but not current, racial economic equality. *Personality and Social Psychology Bulletin*, OnlineFirst. DOI: <https://doi.org/10.1177/0146167220942625>.
- Onyeador, I. N., Hudson, S.-k. T., and Lewis Jr, N. A. (2020) Moving Beyond Implicit Bias Training: Policy Insights for Increasing Organizational Diversity. *Policy Insights from the Behavioral and Brain Sciences*, Forthcoming.
- Osborne, S., Lau, I., and Britton, A. (2012) *Homicides, Firearm Offences and Intimate Violence 2010/2011: Supplementary Volume 2 to Crime in England and Wales 2010/11*. London: Home Office.

- Otoyo, E. (2018) Policing of ethnic minorities in Britain. PhD thesis. London Metropolitan University.
- Pashler, H. and Wagenmakers, E.-J. (2012) Editors' introduction to the special section on replicability in psychological science: A crisis of confidence?. *Perspectives on psychological science*, 7 (6), 528–530.
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*. Ed. by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett. Vol. 32. Curran Associates, Inc., 8026–8037.
- Pate, A., Hamilton, E. E., and Annan, S. O. (1992) *Metro-Dade Spouse Abuse Replication Project: Technical Report*. Washington: Police Foundation.
- Pearson, K. (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, 60, 489–498.
- Phillips, C. and Bowling, B. (2007) Racism, ethnicity, crime and criminal justice. *Oxford Handbook of Criminology*. 4th ed. Oxford: Oxford University Press, 579–619.
- Phipps, A. (2013) Violence against sex workers in the UK. *Violence against women: Current theory and practice in domestic abuse, sexual violence and exploitation*. Ed. by N. Lombard and L. McMillan. London: Jessica Kingsley Publishers, 87–102.
- Pierson, E., Simoiu, C., Overgoor, J., Corbett-Davies, S., Jenson, D., Shoemaker, A., et al. (2020) A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour*, 4 (7), 736–745.
- Pope, D. G., Price, J., and Wolfers, J. (2018) Awareness reduces racial bias. *Management Science*, 64 (11), 4988–4995.
- Pruitt, D. G. and Kimmel, M. J. (1977) Twenty years of experimental gaming: Critique, synthesis, and suggestions for the future. *Annual Review of Psychology*, 28 (1), 363–392.
- Quinton, P. (2011) The formation of suspicions: Police stop and search practices in England and Wales. *Policing and Society*, 21 (4), 357–368.
- (2015) Race disproportionality and officer decision-making. *Stop and Search: The Anatomy of a Police Power*. Ed. by M. Delsol Rebekah and Shiner. Basingstoke: Palgrave Macmillan, 57–78.

- Quinton, P., Bland, N., and Miller, J. (2000) *Police stops, decision-making and practice*. Police Research Series Paper 130. Home Office, Policing & Reducing Crime Unit.
- Quispe-Torreblanca, E. G. and Stewart, N. (2019) Causal peer effects in police misconduct. *Nature Human Behaviour*, 3 (8), 797–807.
- Quispe-Torreblanca, E. G., Stewart, N., Gathergood, J., and Loewenstein, G. (2019) The red, the black, and the plastic: paying down credit card debt for hotels, not sofas. *Management Science*, 65 (11), 5392–5410.
- Rajah, V., Frye, V., and Haviland, M. (2006) “Aren’t I a victim?” Notes on identity challenges relating to police action in a mandatory arrest jurisdiction. *Violence Against Women*, 12 (10), 897–916.
- Ratcliffe, J. H. and Rengert, G. F. (2008) Near-repeat patterns in Philadelphia shootings. *Security Journal*, 21 (1-2), 58–76.
- Regan, L., Kelly, L., Morris, A., and Dibb, R. (2007) *‘If only we’d known’: An exploratory study of seven intimate partner homicides in Engleshire*. Final Report to the Engleshire Domestic Violence Homicide Review Group. Child and Woman Abuse Studies Unit.
- Reiner, R. (2010) *The politics of the police*. Oxford: Oxford University Press.
- Reinhart, A. (2018) A review of self-exciting spatio-temporal point processes and their applications. *Statistical Science*, 33 (3), 299–318.
- Reinhart, A. and Greenhouse, J. (2018) Self-exciting point processes with spatial covariates: modeling the dynamics of crime. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 67 (5), 1305–1329.
- Richardson, R., Schultz, J., and Crawford, K. (2019) Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. *New York University Law Review Online*, 94, 15–55.
- Ridgeway, G. (2007) *Analysis of racial disparities in the New York Police Department’s stop, question, and frisk practices*. Technical Report. Rand Corporation.
- Ridgeway, G. and MacDonald, J. M. (2009) Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association*, 104 (486), 661–668.
- Rinke, J. and Traxler, C. (2011) Enforcement spillovers. *Review of Economics and Statistics*, 93 (4), 1224–1234.
- Robinson, A. L., Pinchevsky, G. M., and Guthrie, J. A. (2018) A small constellation: risk factors informing police perceptions of domestic abuse. *Policing and Society*, 28 (2), 189–204.

- Roch, A., Ritchie, G., and Morton, J. (2010) *Transgender people's experiences of domestic abuse*. Edinburgh: LGBT Youth Scotland and the Equality Network, Scotland.
- Rogers, M. et al. (2016) Barriers to help-seeking: older women's experiences of domestic violence and abuse—Briefing note. Available from: <https://usir.salford.ac.uk/id/eprint/41328/1/Barriers%20to%20older%20women%27s%20help-seeking%20Briefing%20note%2014%207%202016.pdf>. [Accessed 2021-02-10].
- Rojek, J., Rosenfeld, R., and Decker, S. (2012) Policing race: The racial stratification of searches in police traffic stops. *Criminology*, 50 (4), 993–1024.
- Ross, C. T. (2015) A multi-level Bayesian analysis of racial bias in police shootings at the county-level in the United States, 2011–2014. *PLoS One*, 10 (11), e0141854.
- Ross, C. T., Winterhalder, B., and McElreath, R. (2018) Resolution of apparent paradoxes in the race-specific frequency of use-of-force by police. *Palgrave Communications*, 4 (61).
- Roth, J. and Sant'Anna, P. H. (2021) Efficient Estimation for Staggered Rollout Designs. [Preprint] Available from: <https://arxiv.org/abs/2102.01291>.
- Rotton, J. and Cohn, E. (2001) Temperature, Routine Activities, and Domestic Violence: A Reanalysis. *Violence and Victims*, 16 (2), 203–15.
- Rubin, D. B. (1981) The Bayesian bootstrap. *The Annals of Statistics*, 9 (1), 130–134.
- Rudovsky, D. (2001) Law enforcement by stereotypes and serendipity: Racial profiling and stops and searches without cause. *University of Pennsylvania Journal of Constitutional Law*, 3, 296–366.
- SafeLives (2015) *Insights Idva National Dataset 2013-14*. Available from: <https://safelives.org.uk/sites/default/files/resources/Insights%20Idva%20national%20dataset%202013-2014.pdf>. [Accessed 2021-02-10]. Bristol: SafeLives.
- Sampson, R. J. and Lauritsen, J. L. (1997) Racial and ethnic disparities in crime and criminal justice in the United States. *Crime and Justice*, 21, 311–374.
- Sanders, A. N., Kuhns, J. B., and Blevins, K. R. (2017) Exploring and understanding differences between deliberate and impulsive male and female burglars. *Crime & Delinquency*, 63 (12), 1547–1571.
- Sanders-McDonagh, E., Neville, L., and Nolas, S.-M. (2016) From pillar to post: understanding the victimisation of women and children who experience domestic violence in an age of austerity. *Feminist review*, 112 (1), 60–76.

- Scarman, J. (1981) *The Brixton Disorders, 10–12th April (1981)*. London: Her Majesty’s Stationary Office.
- Scheel, A. M., Tiokhin, L., Isager, P. M., and Lakens, D. (2020) Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, OnlineFirst. DOI: <https://doi.org/10.1177/1745691620966795>.
- Schoenberg, F. P. (2002) On rescaled Poisson processes and the Brownian bridge. *Annals of the Institute of Statistical Mathematics*, 54 (2), 445–457.
- Schuster, E. F. (1985) Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics—Theory and Methods*, 14 (5), 1123–1136.
- Sekhon, N. (2018) Dangerous Warrants. *Washington Law Review*, 93 (2), 967–1017.
- Sharp, D. and Atherton, S. (2007) To serve and protect? The experiences of policing in the community of young people from black and other ethnic minority groups. *The British Journal of Criminology*, 47 (5), 746–763.
- Shiner, M. (2010) Post-Lawrence policing in England and Wales: Guilt, innocence and the defence of organizational ego. *The British Journal of Criminology*, 50 (5), 935–953.
- (2015) Regulation and Reform. *Stop and Search*. Ed. by R. Delsol and M. Shiner. Basingstoke: Palgrave Macmillan, 146–169.
- Shiner, M., Carre, Z., Delsol, R., and Eastwood, N. (2018) *The colour of injustice: ‘Race’, drugs and law enforcement in England and Wales*. London: StopWatch & Release.
- Shiner, M. and Delsol, R. (2015) The politics of the powers. *Stop and Search: The Anatomy of a Police Power*. Ed. by R. Delsol and M. Shiner. Basingstoke: Palgrave Macmillan, 31–56.
- Shiner, M. and Thornbury, P. (2019) *An Evaluation of the Northamptonshire Police Reasonable Grounds Panel: Regulating Police Stop and Search*. London: Open Society Justice Initiative.
- Short, M. B., D’orsogna, M. R., Brantingham, P. J., and Tita, G. E. (2009) Measuring and modeling repeat and near-repeat burglary effects. *Journal of Quantitative Criminology*, 25 (3), 325–339.
- Sileshi, G., Hailu, G., and Nyadzi, G. I. (2009) Traditional occupancy-abundance models are inadequate for zero-inflated ecological count data. *Ecological Modelling*, 220 (15), 1764–1775.
- Sim, J. J., Correll, J., and Sadler, M. S. (2013) Understanding police and expert performance: When training attenuates (vs. exacerbates) stereotypic bias in

- the decision to shoot. *Personality and Social Psychology Bulletin*, 39 (3), 291–304.
- Simmons, J. P., Nelson, L. D., and Simonsohn, U. (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22 (11), 1359–1366.
- Simoiu, C., Corbett-Davies, S., and Goel, S. (2017) The problem of infra-marginality in outcome tests for discrimination. *The Annals of Applied Statistics*, 11 (3), 1193–1216.
- Simpson, D. G. (1987) Minimum Hellinger distance estimation for the analysis of count data. *Journal of the American Statistical Association*, 82 (399), 802–807.
- Simpson, E. H. (1951) The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 13 (2), 238–241.
- Skogan, W. G. (1984) Reporting crimes to the police: The status of world research. *Journal of Research in Crime and Delinquency*, 21 (2), 113–137.
- (2006) Asymmetry in the impact of encounters with police. *Policing & Society*, 16 (2), 99–126.
- Smith, D. J. and Gray, J. (1985) *Police and people in London: The PSI report*. London: Gower Publishing.
- Smith, R. J. (2015) Reducing Racially Disparate Policing Outcomes: Is Implicit Bias Training the Answer?. *University of Hawai'i Law Review*, 37, 295–312.
- Smith, S. J. (1986) *Crime, space and society*. Cambridge: Cambridge University Press.
- Snedecor, G. W. (1946) *Statistical methods*. Ames, Iowa: Iowa State College Press.
- Sobel, M. E. (2006) What do randomized studies of housing mobility demonstrate? Causal inference in the face of interference. *Journal of the American Statistical Association*, 101 (476), 1398–1407.
- Stan Development Team (2018) PyStan: the Python interface to Stan. Version: 2.19.1.1.
- (2020) RStan: the R interface to Stan. Version: 2.19.3.
- Stanford Open Policing Project (2021) Stanford Open Policing Project. <https://openpolicing.stanford.edu/>.
- Stark, L., Warner, A., Lehmann, H., Boothby, N., and Ager, A. (2013) Measuring the incidence and reporting of violence against women and girls in Liberia using the ‘neighborhood method’. *Conflict and Health*, 7 (1), 20.

- Starmer, K. (2011) Domestic violence: The facts, the issues, the future. *International Review of Law, Computers & Technology*, 25 (1-2), 9–15.
- Strube, M. J. (1988) The decision to leave an abusive relationship: empirical evidence and theoretical issues.. *Psychological Bulletin*, 104 (2), 236.
- Tanner, J. (1949) Fallacy of per-weight and per-surface area standard, and their relation to spurious correlation. *Journal of Applied Physiology*, 2, 1–16.
- Taylor, S. C. and Gassner, L. (2010) Stemming the flow: challenges for policing adult sexual assault with regard to attrition rates and under-reporting of sexual offences. *Police Practice and Research: An International Journal*, 11 (3), 240–255.
- Teeselink, B. K., van Loon, R. J. P., van den Assem, M. J., and van Dolder, D. (2020) Incentives, performance and choking in darts. *Journal of Economic Behavior & Organization*, 169, 38–52.
- Terrill, W. (2001) *Police coercion: Application of the force continuum*. New York: LFB Scholarly Publishing.
- The Centre for Social Justice (2018) *It can be stopped—A proven blueprint to stop violence and tackle gang and related offending in London and beyond*. London: The Centre for Social Justice.
- Tiratelli, M., Quinton, P., and Bradford, B. (2018) Does stop and search deter crime? Evidence from ten years of London-wide data. *The British Journal of Criminology*, 58 (5), 1212–1231.
- Tjaden, P. and Thoennes, N. (2000a) Prevalence and consequences of male-to-female and female-to-male intimate partner violence as measured by the National Violence Against Women Survey. *Violence Against Women*, 6 (2), 142–161.
- Tjaden, P. G. and Thoennes, N. (2000b) *Extent, nature, and consequences of intimate partner violence: Findings from the National Violence Against Women Survey*. Washington: National Institute of Justice.
- Trendl, A., Stewart, N., and Mullet, T. (2021) The role of alcohol in the link between national football (soccer) tournaments and domestic abuse - Evidence from England. *Social Science & Medicine*, 268, 113457.
- Tressoldi, P. E. (2012) Replication unreliability in psychology: elusive phenomena or “elusive” statistical power?. *Frontiers in Psychology*, 3, 218.
- Trute, B., Adkins, E., and MacDonald, G. (1992) Professional attitudes regarding the sexual abuse of children: Comparing police, child welfare and community mental health. *Child Abuse & Neglect*, 16 (3), 359–368.

- Tur-Prats, A. (2019) Family Types and Intimate Partner Violence: A Historical Perspective. *Review of Economics and Statistics*, 101 (5), 878–891.
- Tyler, T. R., Fagan, J., and Geller, A. (2014) Street stops and police legitimacy: Teachable moments in young urban men’s legal socialization. *Journal of Empirical Legal Studies*, 11 (4), 751–785.
- Waddington, P. A. (1999) Police (canteen) sub-culture. An appreciation. *The British Journal of Criminology*, 39 (2), 287–309.
- Waddington, P. A., Stenson, K., and Don, D. (2004) In Proportion: Race, and Police Stop and Search. *British Journal of Criminology*, 44 (6), 889–914.
- Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., and Kievit, R. A. (2012) An agenda for purely confirmatory research. *Perspectives on Psychological Science*, 7 (6), 632–638.
- Walby, S. and Allen, J. (2004) *Domestic violence, sexual assault and stalking: Findings from the British Crime Survey*. Home Office Research Study 276. Home Office Research, Development and Statistics Directorate.
- Walby, S. and Towers, J. (2012) Measuring the impact of cuts in public expenditure on the provision of services to prevent violence against women and girls. Report for Northern Rock Foundation and Trust for London.
- Walklate, S. (2008) What is to be done about violence against women? Gender, violence, cosmopolitanism and the law. *The British Journal of Criminology*, 48 (1), 39–54.
- Ward, L., Nicholas, S., and Willoughby, M. (2011) *An assessment of the Tackling Knives and Serious Youth Violence Action Programme (TKAP)—phase II*. Research Report 53. Home Office.
- Warren, P., Tomaskovic-Devey, D., Smith, W., Zingraff, M., and Mason, M. (2006) Driving while black: Bias processes and racial disparity in police stops. *Criminology*, 44 (3), 709–738.
- Weisburd, D. (2015) The law of crime concentration and the criminology of place. *Criminology*, 53 (2), 133–157.
- Weisburd, D., Wooditch, A., Weisburd, S., and Yang, S.-M. (2016) Do stop, question, and frisk practices deter crime? Evidence at microunits of space and time. *Criminology & Public Policy*, 15 (1), 31–56.
- Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S., and Weinberger, M. J. (2003) *Inequalities for the L1 deviation of the empirical distribution*. Technical Report HPL-2003-97R1. Hewlett-Packard Labs.

- Westmarland, N. and Kelly, L. (2013) Why extending measurements of ‘success’ in domestic violence perpetrator programmes matters for social work. *British Journal of Social Work*, 43 (6), 1092–1110.
- Whitfield, J. (2004) *Unhappy dialogue. The Metropolitan Police and Black Londoners in postwar Britain*. Cullompton: Willan Publishing.
- Williams, P. (2018) *Being Matrixed: the (over) policing of gang suspects in London*. London: StopWatch.
- Wilson, I. M., Graham, K., and Taft, A. (2017) Living the cycle of drinking and violence: A qualitative study of women’s experience of alcohol-related intimate partner violence. *Drug and Alcohol Review*, 36 (1), 115–124.
- Winking, J. and Mizer, N. (2013) Natural-field dictator game shows no altruistic giving. *Evolution and Human Behavior*, 34 (4), 288–293.
- Withrow, B. L. and Williams, H. (2015) Proposing a benchmark based on vehicle collision data in racial profiling research. *Criminal Justice Review*, 40 (4), 449–469.
- Wolfowitz, J. (1957) The minimum distance method. *The Annals of Mathematical Statistics*, 28 (1), 75–88.
- Women’s Aid (2009) *The Survivor’s Handbook*. Available from: <https://www.womensaid.org.uk/wp-content/uploads/2016/05/Full-Survivors-Handbook-English-2009.pdf>. [Accessed 2021-02-10]. Britol: Women’s Aid.
- Yatracos, Y. G. (1985) Rates of convergence of minimum distance estimators and Kolmogorov’s entropy. *The Annals of Statistics*, 13 (2), 768–774.
- Yeo, I.-K. and Johnson, R. A. (2001) A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. *Statistics & Probability Letters*, 51 (1), 63–69.
- Zhuang, J. (2006) Second-order residual analysis of spatiotemporal point processes and applications in model evaluation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68 (4), 635–653.
- (2011) Next-day earthquake forecasts for the Japan region generated by the ETAS model. *Earth, Planets and Space*, 63 (3), 207–216.
- (2020) Estimation, diagnostics, and extensions of nonparametric Hawkes processes with kernel functions. *Japanese Journal of Statistics and Data Science*, 3 (1), 391–412.
- Zhuang, J. and Mateu, J. (2019) A semiparametric spatiotemporal Hawkes-type point process model with periodic background for crime data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 182 (3), 919–942.

Zhuang, J., Ogata, Y., and Vere-Jones, D. (2002) Stochastic declustering of space-time earthquake occurrences. *Journal of the American Statistical Association*, 97 (458), 369–380.